#6

Docket No. 826.1669/JDH

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:

Hiroshi TSUDA

Serial No.: To be assigned

Filed: January 23, 2001

Group Art Unit: Unassigned

Examiner: Unassigned

For: DOCUMENT SEARCHING APPARATUS, METHOD THEREOF, AND RECORD MEDIUM THEREOF

SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN APPLICATION IN ACCORDANCE WITH THE REQUIREMENTS OF 37 C.F.R. §1.55

Assistant Commissioner for Patents
Washington, D.C. 2023l

Sir:

In accordance with the provisions of 37 C.F.R. §1.55, the applicant submits herewith a certified copy of the following foreign application:

Japanese Patent Application No. 2000-028299
Filed: February 4, 2000.

It is respectfully requested that the applicant be given the benefit of the foreign filing date as evidenced by the certified papers attached hereto, in accordance with the requirements of 35 U.S.C. §119.

Respectfully submitted,

STAAS & HALSEY LLP

Date: January 23, 2001

By: _____

James D. Halsey, Jr.
Registration No. 22,729

700 Eleventh Street, N.W.
Suite 500
Washington, D.C. 20001
(202) 434-1500

#6

PATENT OFFICE

JAPANESE GOVERNMENT

This is to certify that the annexed is a true copy of the following application as filed with this office.

Date of Application: February 4, 2000

Application Number: Patent Application No. 2000-028299

Applicant(s): FUJITSU LIMITED

September 29, 2000

Commissioner,

Patent Office Kozo Oikawa

Certificate No. 2000-3079667

日 本 国 特 許 庁
PATENT OFFICE
JAPANESE GOVERNMENT

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日
Date of Application: ２０００年 ２月 ４日

出 願 番 号
Application Number: 特願２０００－０２８２９９

出 願 人
Applicant (s): 富士通株式会社

CERTIFIED COPY OF
PRIORITY DOCUMENT

２０００年 ９月２９日

特許庁長官
Commissioner,
Patent Office 及川耕造

出証番号 出証特２０００－３０７９６６７

【書類名】　特許願
【整理番号】　0050081
【提出日】　平成12年 2月 4日
【あて先】　特許庁長官殿
【国際特許分類】　G06F 17/30
【発明の名称】　文書検索装置及び方法並びに記録媒体
【請求項の数】　31
【発明者】
　【住所又は居所】　神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
　【氏名】　津田　宏
【特許出願人】
　【識別番号】　000005223
　【氏名又は名称】　富士通株式会社
【代理人】
　【識別番号】　100074099
　【住所又は居所】　東京都千代田区二番町８番地２０　二番町ビル３Ｆ
　【弁理士】
　【氏名又は名称】　大菅　義之
　【電話番号】　03-3238-0031
【選任した代理人】
　【識別番号】　100067987
　【住所又は居所】　神奈川県横浜市鶴見区北寺尾７－２５－２８－５０３
　【弁理士】
　【氏名又は名称】　久木元　彰
　【電話番号】　045-573-3683
【手数料の表示】
　【予納台帳番号】　012542
　【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9705047

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　文書検索装置及び方法並びに記録媒体

【特許請求の範囲】

【請求項１】　リンク関係を有する文書データ群から文書データを検索する文書検索装置であって、

前記リンク関係に重みをつけて重要度であるリンク重要度を前記文書データに付与するリンク重要度付与手段と、

前記リンク重要度に基づいて前記文書データにアクセスするアクセス手段と、

を備えることを特徴とする文書検索装置。

【請求項２】　前記リンク重要度付与手段は、前記文書データを指し示すＵＲＬ（Ｕｎｉｆｏｒｍ　Ｒｅｓｏｕｒｃｅ　Ｌｏｃａｔｏｒ）間の類似度であるＵＲＬ類似度を計算するＵＲＬ類似度計算手段を備え、

前記ＵＲＬ間の類似度と文書データ間の前記リンク関係を用いて前記リンク重要度を計算する、

ことを特徴とする請求項１に記載の文書検索装置。

【請求項３】　前記文書データ中からキーワードを抽出するキーワード抽出手段を更に備える、　ことを特徴とする請求項１または２記載の文書検索装置。

【請求項４】　前記キーワード抽出手段は、前記キーワードの前記文書データ内での出現頻度を算出し、

前記リンク重要度及び前記キーワードの出現頻度を用いて前記キーワードと前記文書データとの関連度を計算するキーワード文書関連度計算手段を更に備える、

ことを特徴とする請求項１乃至３のいずれかに記載の文書検索装置。

【請求項５】　利用者からのアクセスを監視し、アクセスログを作成する監視手段をさらに備え、

前記キーワード文書関連度計算手段は、前記キーワード出現頻度、前記リンク重要度及び前記アクセスログを用いて、関連度を計算する、

ことを特徴とする請求項４に記載の文書検索装置。

【請求項６】　前記ＵＲＬ類似度並びに前記文書データのリンク数及び被リ

ンク数を用いて前記文書データの文書タイプを判別する文書タイプ判別手段をさらに備え、

前記キーワード文書関連度計算手段は、前記文書タイプに基づいて前記文書データを選択し、前記選択された文書データについて前記関連度を計算する、

ことを特徴とする請求項４又は５記載の文書検索装置。

【請求項７】　前記抽出されたキーワードの読みや綴りから前記キーワードへアクセスできる索引を生成する索引生成手段を更に備える、

ことを特徴とする請求項３乃至６いずれか記載の文書検索装置。

【請求項８】　利用者が前記キーワードの読みや綴りの断片を、前記索引から選択する選択手段を更に備え、

前記索引生成手段は前記キーワード文書関連度算出手段による前記関連度の高い一定個数以内の文書データを索引に入れ、

前記アクセス手段は、前記選択された前記キーワードの読みや綴りの断片に基づいて、前記文書データにアクセスする、

ことを特徴とする請求項７記載の文書検索装置。

【請求項９】　ネットワークから前記文書データを収集する収集手段を更に備えることを特徴とする請求項１乃至８いずれか記載の文書検索装置。

【請求項１０】　前記リンク重要度付与手段は類似度の低いＵＲＬから多くリンクされている文書データの重要度を低くすることを特徴とする請求項１記載の文書検索装置。

【請求項１１】　前記リンク重要度付与手段は、重要な文書データからリンクされている、ＵＲＬの類似度の低いページは重要とすることを特徴とする請求項１記載の文書検索装置。

【請求項１２】　前記ＵＲＬの類似度は、サーバアドレス、パス、ファイル名を含むＵＲＬの字面情報に基づいて決定される、

ことを特徴とする請求項２記載の文書検索装置。

【請求項１３】　対象となる文書データ集合をＤＯＣ＝｛p_1，p_2，・・・，p_N｝

文書データｐの重要度をＷｐ

文書データｐのリンク先のページ集合を$R_{ef(p)}$

文書データｐのリンク元のページ集合を$R_{efed(p)}$

文書データｐとｑのＵＲＬ上の類似度をsim(p,q)，相異度をdiff(p,q)=1/sim(p,q)とし、

文書データｐからｑにリンクが張られているとした時、そのリンクの重みlw(p,q)

を以下で定義し、

【数１】

$$lw(p,q) = diff(p,q) / \sum_{i \in Ref(p)} diff(p,i)$$

$$= \frac{1}{sim(p,q) \sum_{i \in Ref(p)} \frac{1}{sim(p,i)}} \qquad \cdots\cdots\cdots(1)$$

各ページの重要度は、各ｐ∈ＤＯＣに対して、C_q を定数（重要度の下限として

【数２】

$$Wq = Cq + \sum_{p \in Refed(q)} W_p * lw(p,q) \qquad \cdots\cdots\cdots(2)$$

という連立一次方程式の解として求めることを特徴とする請求項１記載の文書検索装置。

【請求項１４】　前記文書データはＷＷＷ（Ｗｏｒｌｄ　Ｗｉｄｅ　Ｗｅｂ）ページであることを特徴とする請求項１乃至１３いずれか記載の文書検索装置。

【請求項１５】　リンク関係を有する文書データ群の索引を生成する文書索引生成装置であって、

前記リンク関係を用いて前記文書データにリンク重要度を付与するリンク重要度付与手段と、

前記文書データ中からキーワードを抽出するキーワード抽出手段と、

前記抽出されたキーワードの読みや綴りから前記キーワードへアクセスできる索引を生成する索引生成手段と、

前記索引からキーワードの前記読みや綴りを選択された場合に、キーワード及び前記キーワードに関係する高いリンク重要度を付与された文書データにアクセスするアクセス手段と、

を備えることを特徴とする文書索引生成装置。

【請求項１６】　前記リンク重要度付与手段は、前記文書データを指し示すＵＲＬ（Ｕｎｉｆｏｒｍ　Ｒｅｓｏｕｒｃｅ　Ｌｏｃａｔｏｒ）間の類似度であるＵＲＬ類似度を計算するＵＲＬ類似度計算手段を備え、

前記リンク重要度付与手段は、前記ＵＲＬ間の類似度と文書データ間のリンク関係を用いて前記リンク重要度を計算する、

ことを特徴とする請求項１５記載の文書索引生成装置。

【請求項１７】　リンク関係を有する文書データ群の索引を生成する文書索引生成装置であって、

前記文書データのＵＲＬが類似しているか否かに基づいて前記文書データにリンク重要度を付与するリンク重要度付与手段と、

前記文書データ中からキーワードを抽出するキーワード抽出手段と、

前記抽出されたキーワードの読みや綴りから前記キーワードへアクセスできる索引を前記リンク重要度に基づいて生成する索引生成手段と、

を備えることを特徴とする文書索引生成装置。

【請求項１８】　リンク関係を有する文書データ群へのリンク集を生成するリンク集生成システムであって、

ネットワークから前記文書データを収集する収集手段と、

前記リンク関係を用いて算出された重要度であるリンク重要度を前記文書データに付与するリンク重要度付与手段と、

前記文書データ群から特定の文字列上の特徴を持つＵＲＬを判別するＵＲＬ文

字列判別手段と、

前記リンク重要度及び前記ＵＲＬの特定の文字列上の特徴に基づいて、前記文書データへのリンクを順に一定個数以内並べたリンク集を生成する索引生成手段と、

を備えることを特徴とするリンク集生成システム。

【請求項１９】　前記ＵＲＬ類似度並びに前記文書データのリンク数及び被リンク数を用いて前記文書データの文書タイプを判別する文書タイプ判別手段をさらに備え、

前記索引生成手段は、前記文書タイプに基づいて前記文書データを選択し、前記選択された文書データについて前記リンク重要度及び前記ＵＲＬの特定の文字列上の特徴に基づいて、前記リンク集を生成する、

ことを特徴とする請求項１８に記載のリンク集生成システム。

【請求項２０】　リンク関係を有する文書データ群から文書データを検索する文書検索方法であって、

前記リンク関係を用いて算出した重要度であるリンク重要度を前記文書データに付与する過程と、　前記リンク重要度に基づいて前記文書データにアクセスする過程と、

を含むことを特徴とする文書検索方法。

【請求項２１】　前記文書データを指し示すＵＲＬ（Ｕｎｉｆｏｒｍ　Ｒｅｓｏｕｒｃｅ　Ｌｏｃａｔｏｒ）間の類似度であるＵＲＬ類似度を計算する過程と、

前記ＵＲＬ類似度と文書間のリンク関係を用いて前記リンク重要度を計算する過程と、

を更に含むことを特徴とする請求項２０に記載の文書検索方法。

【請求項２２】　　前記文書データ中からキーワードを抽出する過程、

を更に含むことを特徴とする請求項２０又は２１記載の文書検索方法。

【請求項２３】　前記キーワードの前記文書データ内での出現頻度を算出する過程と、

前記リンク重要度及び前記キーワードの出現頻度を用いて前記キーワードと前

記文書データとの関連度を計算する過程と、

を更に含むことを特徴とする請求項２０乃至２２いずれかに記載の文書検索方法。

【請求項２４】　利用者からのアクセスを監視し、アクセスログを作成する過程と、

前記キーワード出現頻度、前記リンク重要度及び前記アクセスログを用いて、関連度を計算する過程と、

を更に含むことを特徴とする請求項２３に記載の文書検索方法。

【請求項２５】　前記ＵＲＬ類似度並びに前記文書データのリンク数及び被リンク数を用いて前記文書データの文書タイプを判別する過程と、

前記文書タイプに基づいて前記文書データを選択し、前記選択された文書データについて前記関連度を計算する過程と、

を更に含むことを特徴とする請求項２３又は２４記載の文書検索方法。

【請求項２６】　前記抽出されたキーワードの読みや綴りから前記キーワードへアクセスできる索引を生成する過程、

を更に含むことを特徴とする請求項２２乃至２５いずれか記載の文書検索方法。

【請求項２７】　利用者が前記キーワードの読みや綴りの断片を、前記索引から選択する過程と、

前記関連度の高い一定数以下の文書データを前記索引に入れる過程と、

前記選択された前記キーワードの読みや綴りの断片に基づいて、前記文書データにアクセスする過程と、

を更に含むことを特徴とする請求項２６記載の文書検索方法。

【請求項２８】　ネットワークから前記文書データを収集する過程を更に含むことを特徴とする請求項２０乃至２７いずれか記載の文書検索方法。

【請求項２９】　リンク関係を有する文書データ群へのリンク集を自動生成する自動リンク集生成方法であって、

ネットワークから前記文書データを収集する過程と、

前記リンク関係を用いて前記文書データにリンク重要度を付与する過程と、

前記文書データ群から特定の文字列上の特徴を持つＵＲＬを判別する過程と、

前記リンク重要度及び前記ＵＲＬの特定の文字列上の特徴に基づいて、前記文書データへのリンクを順に一定個数以内並べたリンク集を生成する過程と、

を含むことを特徴とするリンク集生成方法。

【請求項３０】　前記ＵＲＬ類似度並びに前記文書データのリンク数及び被リンク数を用いて前記文書データの文書タイプを判別する過程と、

前記文書タイプに基づいて前記文書データを選択し、前記選択された文書データについて前記リンク重要度及び前記ＵＲＬの特定の文字列上の特徴に基づいて、前記リンク集を生成する過程と、

を更に含むことを特徴とする請求項２９に記載のリンク集生成方法。

【請求項３１】　リンク関係を有する文書データ群へのリンク集を生成するコンピュータのためのプログラムを記録した記録媒体であって、

ネットワークから前記文書データを収集するステップと、

前記リンク関係を用いて前記文書データにリンク重要度を付与するステップと、

前記文書データ群から特定の文字列上の特徴を持つＵＲＬを判別するステップと、

前記リンク重要度及び前記ＵＲＬの特定の文字列上の特徴に基づいて、前記文書データへのリンクを順に一定個数以内並べたリンク集を生成するステップと、

を含む処理を前記コンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、情報処理装置に蓄えられた大量の文書ファイル群を、文書の内容、文書のリンク関係及び文書の登録された場所等に基づいて検索する文書検索装置及びその方法並びに記録媒体に関する。

【０００２】

【従来の技術】

今日、コンピュータネットワークの発達により、多量のオンライン文書情報があふれてきている。これらの大量のオンライン文書情報を検索及び整理するために、オンライン文書情報に索引をつけるサービスが提供されている。

【０００３】

例えば、インターネットのＷＷＷ（Ｗｏｒｌｄ　Ｗｉｄｅ　Ｗｅｂ）ページ検索において、ある有名なサービスでは、ディレクトリサービスを提供している。これは、階層化されたカテゴリ毎に、ＷＷＷページへのリンク集をまとめたものであり、以下の利点がある。

１．カテゴリを選択する（クリックする）だけで、利用者が閲覧を希望するＷＷＷページへのリンクが得られる。

２．ＷＷＷページがカテゴリ毎にまとめられているため、文書検索の際に無闇に多くの情報が検索されない。

３．人手を介してＷＷＷページをカテゴリに登録しているので、ゴミ、つまり、利用者が閲覧を希望している情報と関係のない情報が紛れ込むことが少ない。

【０００４】

これら利点のため、このサービスはインターネットにおいて非常に広く利用されている。しかし、このサービスは人手によりカテゴリの作成やＷＷＷページの管理をしているため、運用コストがかかるという問題点がある。

【０００５】

このディレクトリサービス全体の自動化にあたっては、以下の解決すべき課題がある。

１．重要文書の選定

２．カテゴリ階層の管理、例えば、トピックの追加、削除

３．文書からカテゴリの割付：自動分類

ここで、重要文書の選定について説明する。インターネット、イントラネット共に、ＷＷＷページは増加の一途にある。従って、似たような情報があちこちで別々に作られたりもするので、文書に含まれる文字列（キーワード）の有無や多少により検索しても結果が多すぎることになる。延いては、この多くの検索結果のうちどの情報が重要なのか分からないという問題が生じる。この重要文書の選

定に対する解決手段として以下のようなものが取られてきている。

１．検索要求を満たす順に検索結果を並び換えする、つまり、文書中に含まれる検索キーワードの数などに基づいて、検索結果をソートし、ランキング付けをする。

２．検索結果の視覚化によるアクセス支援を行う、つまり、検索結果の文書群を、内容の近いもの同士グルーピングする（クラスタリング）。

３．文書の属性、例えば、サイズ、作成日時等に基づいて、検索結果をソートする。

４．その他何らかの手段で付与された文書の重要度順でソートする、例えば、リンク関係、ユーザのアクセスログ解析、第３者機関の作成した格付け等のメタデータに基づいて、検索結果をソートする。

【０００６】

最後の一例として、近年WWWページのようなハイパーテキストのリンク関係を利用した文書重要度付与が、研究、サービスレベルで重要な技術となってきている。ここで、リンク関係に基づくリンク重要度付与の最も単純な実現は、「被リンク数の多い文書は重要度が高い」という直観に基づくものである。

【０００７】

【発明が解決しようとする課題】

しかし、情報のナビゲーションを容易にするために、一般に同一サーバ内のWWWページは、互いにリンク関係にあることが多い。例えば、個人WWWページなどでは、「ＸＸのトップに戻る」といった、ホームページトップへのリンクなどが多かったりする。したがって、単純な被リンク数だけでは、大量のページを抱えるサーバ（サイト）や個人が単に量が多いというだけで重要度が高いことになるという問題がある。また、検索システムが被リンク数で重要度を出していることがわかれば、意味のないページ分割や無駄なリンクだけのページを加えることで、故意に自分のページの重要度をあげることができるという問題もある。

【０００８】

また、http://www.elseviernl/cas/tree/store/comnet/free/www7/02/com02.htmにおいて閲覧可能であるWWWページにおいて、

「被リンク数の多い文書は重要度が高い。」
という直感の他に、
「重要度が高い文書からリンクされている文書は重要度が高い。」及び
「リンク先の少ないページからリンクされているページは重要度が高い。」という直感も示唆されている。

【０００９】

二番目の直観は、有名ディレクトリサービスで紹介されるＷＷＷページと、名もない個人のリンク集で紹介されているＷＷＷページとでは、前者の方がより重要という発見に基づいたものである。三番目の直感は、例えば、1000ものリンク先を持つリンク集からリンクされている文書よりも、50のリンク先を持つリンク集からリンクされている文書の方が重要度が高いという考えに基づく。
これらの直観に基づいた重要度判定アルゴリズムにおいて、まず被リンク数により仮重要度を計算し、仮重要度からリンク関係により重要度を更新し、・・・という操作を収束するまで繰り返すことを行っている。

【００１０】

しかし、このアルゴリズムによっても、大量のページを抱えるサイトは被リンク数が多くなるため有利であり、その結果、ページの重要度を算出した場合に重要度の高いページの上位に似たようなサイトのページが並ぶという上述の問題がある。

【００１１】

ところで、利用者が検索をする際、検索に用いるキーワードへアクセスするインターフェースが必要となる。キーワードへのアクセスに関する従来技術として、かな漢字変換のインタフェースと関係が深い。

【００１２】

例えば、特開平０３－２４１４５６によれば、タッチパネル式のデバイスにより、かな漢字変換を実現している。この技術によれば、画面上のソフトウェアキーボードにより、キーワードの読みを全て入れてから、「変換」キーを押して漢字まじりに変換している。

【００１３】

また、例えば、特開平１０－１５４１４４、特開平１０－１５４０３３及びhttp://www.csl.sony.co.jp/person/masui/POBox/index.htmにおいて閲覧可能であるＷＷＷページによれば、ペン向きテキスト入力システムを実現している。この技術によれば、画面上のソフトウェアキーボードから読みを入力するのであるが、読みの一部分を入力した時点で利用者の使用履歴に基づき随時候補を出していく。

【００１４】

特開平０３－２４１４５６、特開平１０－１５４１４４、特開平１０－１５４０３３及び上述のＷＷＷページによれば、かな漢字変換するために常に1文字ずつキーワードの読みや綴りを入力する必要があるため、長い文字列を入力する

また、例えば、次のような自明な読み入力インタフェースもある。この技術の例として、まず、「あ」「い」・・・といった、最初の文字毎に、キーワードリストを作り、そのキーワードリストのなかから利用者が選択する方式がある。しかい、この場合、特定の読みから始まるリスト内のキーワード数が多い場合、利用者がこのキーワードリストから特定のキーワードを選択することが困難になる。この方式の例として、銀行の自動振込機が挙げられる。

【００１５】

自明な読み入力インタフェース技術のさらなる例として、読みの文字を順次入力（ポインティングデバイスの場合、クリック等）して、その読みから始まるキーワード数が決まった所でかな漢字変換された文字を表示する方式もある。図３２は、読みの文字を順次入力して、キーワード数が決まった所でかな漢字変換された文字を表示する方式を示す。図３２では「秋葉原」を表示する場合を例として示す。図３２に示すように、５０音の中から利用者は、キーワードの読みを順次入力する。「秋葉原」を表示させる場合、利用者は、まず「あ」を入力し、次に「き」、「は」、「ば」「ら」と順次入力する。「あきはばら」全ての音の入力が終了してキーワード数が決まったところで、かな漢字変換された文字「秋葉原」が表示される。この方式によっても、長い読みのキーワードに辿りつくまで、多くの音を入力する必要がある。

【００１６】

本発明は、重要文書の選定において、上述の問題、つまり、多くのページを有する特定サイトに重要度が偏るという問題を解消し、悪意を持った特定個人が重要度を操作することを困難にする文書検索装置及び方法を提供することを目的とする。

【００１７】

さらに、本発明は、検索キーワードをキーワードの読みに基づいて入力する場合に、より少ない入力によりキーワードに到達でき、かつ、表示されるキーワードの候補や文書数を一定数以下にすることにより、キーワード及び文書の選択を容易にする文書検索装置及び方法を提供することを目的とする。

【００１８】

さらに、本発明は、ディレクトリサービス風のインターフェースを持ち、キーワードと関連する重要文書、例えばＷＷＷページへ迅速にアクセスできるリンク集を生成する装置及び方法を提供することを目的とする。

【００１９】

【課題を解決するための手段】

本発明によれば、リンク関係を有する文書データ群から文書データを検索する文書検索装置であって、リンク関係に重みを付けた重要度であるリンク重要度を文書データに付与するリンク重要度付与手段と、リンク重要度に基づいて前記文書データにアクセスするアクセス手段を備える。

【００２０】

多くのページからリンクされているページは重要であると考えられること、及び、少ないページにリンクしているページからのリンクは、多くのページにリンクしているページからのリンクよりも、重要である等と考えられることに基づいて、リンク重要度付与手段はリンク関係に重みを付けてリンク重要度を算出し、文書データに付与する。アクセス手段はこの算出されたリンク重要度に基づいて文書データにアクセスする。これにより、重要な文書データを自動的に検索することを可能とする。　さらに、リンク重要度付与手段は、ＵＲＬ類似度計算手段を備えてもよい。ＵＲＬ類似度計算手段は、文書データを指し示すＵＲＬ（Ｕｎｉｆｏｒｍ　Ｒｅｓｏｕｒｃｅ　Ｌｏｃａｔｏｒ）間の類似度であるＵＲＬ類似

度を計算し、リンク重要度付与手段は、ＵＲＬ間の類似度と文書間のリンク関係を用いてリンク重要度を計算して文書データに付与する。

【００２１】

例えば同じサイト内の文書データはリンク関係を有することが多い。これら同じサイト内文書データのＵＲＬは一般にＵＲＬ類似度が高い。ＵＲＬ類似度が高い文書データからのリンクの重みをＵＲＬ類似度が低い文書データからのリンクの重みよりも低くすることにより、多量の文書データを抱えるサイトの重要度が過大に評価されることを防ぐ。延いては、より精度良く、重要な文書を検索することを可能とする。また、リンク重要度を付与する際にＵＲＬ類似度を考慮するため、サイトの中の文書データのリンク数を増やしたりすることにより第三者が特定の文書の重要度を故意に上げようとすることが難しくなるという効果もある。なお、ＵＲＬ類似度は、サーバアドレス、パス、ファイル名を含むＵＲＬの字面情報に基づいて決定されることとしてもよい。

【００２２】

また、文書検索装置は、更に文書データ中からキーワードを抽出するキーワード抽出手段を更に備えてもよい。

また、文書検索装置は、文書関連度計算手段を更に備え、上述のキーワード抽出手段は抽出されたキーワードの文書データ内での出現頻度を算出し、文書関連度計算手段はリンク重要度及びキーワードの出現頻度を用いてキーワードと文書データとの関連度を計算する。

【００２３】

リンク重要度及び文書中のキーワードの出現頻度を用いて関連度を計算し、関連度の高い文書データを検索することにより、利用者が検索したいと望む文書データと内容が関連している可能性の高い、重要な文書を検索することを可能とする。

【００２４】

また、文書検索装置は、利用者からのアクセスを監視し、アクセスログを作成する監視手段をさらに備え、キーワード文書関連度計算手段は、キーワード出現頻度、リンク重要度に加えてアクセスログを用いて関連度を計算することとして

もよい。関連度を算出する際にアクセスログを用いることにより、よりキーワードに関連した、より重要度の高い文書データを検索することが可能となる。

【００２５】

リンク重要度、キーワード出現頻度及びアクセスログを関連度の計算に用いるため、故意に特定の文書データの重要度を上げた場合であっても、そのような文書データは検索されにくくなるという効果もある。

【００２６】

また、文書検索装置は、ＵＲＬ類似度並びに文書データのリンク数及び被リンク数を用いて文書データの文書タイプを判別する文書タイプ判別手段をさらに備え、キーワード文書関連度計算手段は、文書タイプに基づいて文書データを選択し、選択された文書データについて関連度を計算することとしてもよい。

【００２７】

文書データには、リンクページ、コンテンツページ等、数タイプある。これらは、リンク数及び被リンク数によって判別することが可能である。この文書タイプに基づいて、あるタイプの文書データ、例えば、コンテンツページを選択し、選択された文書について関連度を計算することにより、より精度良く検索することが可能となる。

【００２８】

また、文書検索装置は、抽出されたキーワードの読みや綴りからキーワードへアクセスできる索引を生成する索引生成手段を更に備えてもよい。

さらに、文書検索装置は、利用者が前記キーワードの読みや綴りの断片を、前記索引から選択する選択手段を更に備えてもよい。そして、索引生成手段は索引を生成する際に、算出された関連度が高い順に、一定数以内の文書データを検索にいれる。アクセス手段は選択されたキーワードの読みや綴りの断片に基づいて、文書データにアクセスする。利用者は索引の中に含まれるデータが一定数にかぎられているために、索引から容易に選択することが可能となる。また、索引の中に含まれるデータが一定数にかぎられているため、携帯電話のような移動端末においては表示画面の広さが限られている場合に有用である。

【００２９】

また、文書検索装置はネットワークから文書データを収集する収集手段を更に備えてもよい。

さらに、本発明によれば、リンク関係を有するリンク集生成システムにおいて、収集手段、リンク重要度付与手段、ＵＲＬ文字列判別手段及び索引生成手段を備える。

【００３０】

収集手段はネットワークから文書データを収集し、リンク重要度付与手段は文書データのリンク関係を用いて算出されたリンク重要度を文書データに付与し、ＵＲＬ文字列判別手段は文書データから特定の文字列上の特徴を持つＵＲＬを判別する。索引生成手段は、リンク重要度及びＵＲＬの特定の文字列上の特徴に基づいて、文書データへのリンクを順に一定個数以内並べたリンク集を生成する。文書データのＵＲＬの特定の文字列上の特徴は、その文書データの内容を示すことがある。例えば、ＪＡＶＡに関する文書データのＵＲＬには、ＪＡＶＡやｊａｖａといった文字列が用いられることがある。従って、ＵＲＬの特定の文字列上の特徴は、文書データの内容を推定することに用いることができる。故に、リンク重要度及びＵＲＬの特定の文字列上の特徴に基づいて文書データへのリンク集を作成することにより、利用者が閲覧を希望する内容を含む文書データを検索することが可能なリンク集を自動的に生成することが可能となる。

【００３１】

また、リンク集生成システムは、さらに文書タイプ判別手段を備えてもよい。文書タイプ判別手段は、上述のようにして文書データの文書タイプを判別し、索引生成手段は、文書タイプに基づいて文書データを選択し、選択された文書データについてリンク重要度及びＵＲＬの特定の文字列上の特徴に基づいて、リンク集を生成する。これにより、より質の高い文書データへのリンク集を生成することが可能となる。

【００３２】

また、本発明の範囲は、上述の装置が実現する処理の手順からなる方法をも含む。

また、さらに、本発明の範囲は、上述の処理を前記コンピュータに実行させう

るプログラムを記録する記録媒体をも含む。

【００３３】

【発明の実施の形態】

以下、図面を参照しながら本発明の実施の形態を詳細に説明する。

図１は第１実施形態に係わる文書検索装置の構成を示す。図１の文書検索装置は、処理装置１１、入力装置１２及び表示装置１３を備える。処理装置１１は、例えば、ＣＰＵ（Ｃｅｎｔｒａｌ　Ｐｒｏｃｅｓｓｉｎｇ　Ｕｎｉｔ）とメモリを含み、入力装置１２はキーボード、マウス等に対応し、表示装置１３はディスプレイ等に対応する。

【００３４】

また、処理装置１１は、リンク重要度付与器２１、キーワード抽出器２２、キーワード文書関連度計算器２３、索引生成器２４、索引アクセス部２５及びアクセス解析器２６を備える。これらは、例えば、プログラムにより記述されたソフトウェアコンポーネントに対応し、処理装置１１の特定のプログラムコードセグメントに格納される。

【００３５】

リンク重要度付与器２１は、例えばWWWページのような文書データ３０からリンク情報を抽出する。例えば、ＷＷＷページの場合ＨＴＭＬを解析し、 富士通トップのようにアンカー(a) タグの部分を取り出す。そして、抽出されたリンク情報に基づいてリンク重要度３１を算出し、算出されたリンク重要度３１をキーワード文書関連度計算器２３へ出力する。また、リンク重要度付与器２１は、リンク元とリンク先のＵＲＬの字面の類似度であるＵＲＬ類似度を算出するＵＲＬ類似度計算器２７を備える。ＵＲＬ類似度計算器２７はリンク元とリンク先のＵＲＬ類似度を算出し、リンク重要度付与器２１は、抽出されたリンク関係及びＵＲＬ類似度に基づいてリンク重要度３１を算出する。

【００３６】

キーワード抽出器２２は文書データ３０からキーワードを抽出し、その結果をページキーワード３２として出力する。また、キーワード抽出器２２は抽出され

たキーワードが文書データ３０中で出現する全出現頻度を集計することとしてよよい。キーワード抽出器２２は、例えば、文書データ３０が日本語で記述されている場合に、形態素解析（単語切り）を行い、名詞（列）をキーワードとして抽出する。また、簡単な表記の揺れ（「コンピュータ」と「コンピューター」など）はルールや小規模なルールで統一しておく。同義語の情報は例えば、外付けで辞書などにより与えられる。

【００３７】

キーワード文書関連度計算器２３は、リンク重要度３１、ページキーワード３２及び後述のアクセスログ３４に基づいて、キーワードと文書の関連度であるキーワード文書関連度を計算し、計算結果を索引生成器２４へ出力する。

【００３８】

索引生成器２４は、キーワード文書関連度に基づいて索引データ３３を生成し、生成された索引データ３３を索引アクセス部２５に出力する。索引データ３３は例えばハイパーテキストを用いて生成される。

【００３９】

索引アクセス部２５は、入力装置１２から入力される利用者の指示に従って、索引データ３３の内容を表示装置１３に表示し、利用者のアクセス状況を示す情報をアクセス解析器２６に出力する。

【００４０】

アクセス解析器２６は、利用者のアクセス状況を示す情報を解析し、一定期間内に各キーワードからアクセスされた文書を集計したアクセスログ３４を作成し、作成されたアクセスログ３４をキーワード文書関連度計算器２３に出力する。

【００４１】

次に、図２から図５を参照しながら各種データのデータ構造について説明する。

図２は文書情報を格納するテーブル集を示す。文書情報を格納するテーブル集は、文書情報テーブル４１及び被参照文書テーブル４２を含む。文書情報テーブル４１は、文書ＩＤ、ＵＲＬ、タイトル、被参照文書テーブル４２へのリンク及び文書データのリンク重要度からなる。文書ＩＤは文書データにユニークにつけ

られた数値、ＵＲＬはインターネットで文書データを指す情報、タイトルはその文書データのタイトル、被参照文書テーブル４２はその文書データにリンクしているリンク元の文書データの集合を格納する。被参照文書テーブル４２は文書ＩＤ及びリンク先とリンク元の文書データのＵＲＬ上の字面の類似度であるＵＲＬ類似度の対情報を格納する。１文書毎に高々１つの被参照文書テーブルがある。文書情報テーブル４１及び被参照文書テーブル４２は、リンク重要度付与器２１が作成するリンク重要度３１に相当する。

【００４２】

図３はキーワード情報を格納するテーブル集を示す。キーワード情報を格納するテーブル集はキーワードテーブル５１、キーワード対応テーブル５２及び出現文書テーブル５３を含む。キーワードテーブル５１は、キーワードＩＤ、代表語、出現文書テーブル５３へのリンクからなる。代表語とは、同じキーワードＩＤを持つキーワードが複数ある場合に、どれを代表とするかを示すの情報である。キーワード対応テーブル５２は、実際のキーワードと、読み（又はスペル）、キーワードＩＤとの対情報からなる。ここで、同一概念を表すキーワード（例えば、「コンピュータ」「computer」「計算機」など）には同一のキーワードＩＤ（ｋｗＩＤ）がふられている。また、日本語のキーワードの読みについては、読みの中の「ー」を除去したり、「ぁ」や「ぃ」のような拗撥音を「あ」や「い」としたりして表記を統一しておく。英語のキーワードについては、大文字に統一しておく。これにより、「コンピューター」や「コンピュータ」といった表記のゆれによって同じ概念を示すキーワードが異なるように扱われることを防ぎ、生成される索引においてキーワードの統一を行うことができる。出現文書テーブル５３は、当該キーワードを含む文書データの文書ＩＤとそのキーワードの出現頻度からなる。キーワード対応テーブル５２及びキーワードテーブルの代表語は予め処理装置１１に備えられている（不図示）。出現文書テーブル５３はキワード抽出器２２により作成されるページキーワード３２に相当する。

【００４３】

図４は索引情報を格納するテーブル集を示す。索引情報を格納するテーブル集は、索引情報テーブル６１、関連文書テーブル６２及び関連キーワードテーブル

６３を含む。索引情報テーブル６１は、索引文字列、継続する文字列の並び、関連キーワード列からなる。索引情報テーブル６１は、図３のキーワード対応テーブル５２に格納されたキーワードとその読み（又はスペル）及びキーワード文書に基づいて、索引データ生成部２４が後述の方法により文字列グラフを生成し、それを縮退することにより作成される。図４では、例えば、トップが「あ」「い」・・・。「あ」を選ぶと次に「あいぼ」，「あお」，・・・が現れることを示している。また、文字列「あいぼ」から考えられるキーワードは「相棒」「アイボリー」であることを示している。ここでキーワードは、図４のキーワード対応テーブル５２に含まれるキーワードである。関連文書テーブル６２は、キーワードＩＤから関連する文書のＩＤである関連文書ＩＤを得るためのテーブルである。キーワード文書関連度計算器２３は文書関連度を計算し、その計算結果に基づいて、文書関連度の高いものから順に一定個数以内の関連文書ＩＤ列を関連文書テーブル６２に格納する。関連キーワードテーブル６３は、文書ＩＤから関連キーワードＩＤを得るためのテーブルである。関連文書テーブル６２と関連キーワードテーブル６３の情報は同じで、それを転置したものである。なお、関連文書ＩＤの詳細情報は、図２の文書情報テーブル４１に格納されている。

【００４４】

図５は、アクセスログ情報を格納するテーブルであるアクセスログを示す。アクセスログ７１は、利用者がキーワード情報画面（後述）から文書データを選んだ際に関する情報、つまり、アクセス日時、キーワードＩＤ及び選ばれた文書データの文書ＩＤの対情報等を格納する。アクセスログ７１は、アクセス解析器２６が作成するアクセスログ３４に相当する。一定期間内のログを集計することにより、特定キーワードから特定文書がどのくらいアクセスされたかを得ることができる。

【００４５】

以下、文書検索装置全体の動作について図６を用いて説明する。図６は索引生成処理の手順を示す。

まず、リンク重要度付与器２１は、文書データからリンク情報及びＵＲＬ等を抽出する。続いて、抽出された情報を文書情報テーブル４１の文書ＩＤ、ＵＲＬ

フィールドに格納し、被参照文書テーブル４２へのリンク（ポインタ）及び被参照文書テーブル４２を生成する（ステップＳ１）。

【００４６】

リンク重要度付与器２１に備えられたＵＲＬ類似度計算器２７は、抽出されたリンク情報及びＵＲＬに基づいて、リンク元とリンク先のＵＲＬ類似度を計算し、被参照文書テーブル４２のフィールドに格納する。

【００４７】

次に、リンク重要度付与器２１は抽出されたリンク情報及びＵＲＬ類似度に基づいてリンク重要度を算出し、文書情報テーブル４１のフィールドに格納する（ステップＳ２）。ＵＲＬ類似度及びリンク重要度の算出については後述する。

【００４８】

キーワード抽出器２２は、文書データからキーワードを抽出し、キーワード対応テーブル５２のキーワードとキーワードＩＤのフィールド、キーワードテーブル５１の全フィールド及び出現文書テーブル５３の文書ＩＤと頻度のフィールドに格納する（ステップＳ３）。キーワードの抽出は、例えば、文書データ３０が日本語で記述されている場合に、キーワード抽出器２２は形態素解析（単語切り）を行い、単語切りによって得られた名詞（列）から行う。また、簡単な表記の揺れ（「コンピュータ」と「コンピューター」など）はルールや小規模な辞書で統一しておく。同義語の情報は外付けで辞書などにより与えられる。

【００４９】

次にキーワード抽出器２２は、抽出されたキーワードの読み付けを上述のようにして統一された表記法則に基づいて行い、その結果をキーワード対応テーブル５２の読みフィールド（又はスペル）に格納する（ステップＳ４）。キーワード対応テーブル５２ではキーワードの表記が統一されて格納されているため、これにより、生成される索引のキーワードを統一することができる。

【００５０】

次にキーワード抽出器２２は、抽出されたキーワードが文書データ中で出現する全出現頻度を集計し、その結果に基づいてキーワードテーブル５１の出現文書のポインタを生成し、出現文書テーブル５３の文書ＩＤ及び頻度のフィールドに

集計された頻度を格納する（ステップＳ５）。さらに、キーワード抽出器２２は、キーワードＩＤの全出現頻度を集計し、上位一定個数（例えば10,000語）のキーワードを索引の対象キーワードとして決定し、当該キーワードＩＤ以外のエントリを、キーワードテーブル５１及びキーワード対応テーブル５２から除く。

【００５１】

次に、キーワード文書関連度計算器２３は、文書情報テーブル４１のリンク重要度、被参照文書テーブル４２のＵＲＬ類似度及びアクセスログ７１に基づいてキーワードと文書の関連度を示すキーワード文書関連度を計算し、キーワード文書関連度が上位から一定個数にある文書を関連文書として決定し、その決定に基づいて関連文書テーブル６２及び関連キーワードテーブル６３の関連文書ＩＤ列のフィールドに情報を格納する（ステップＳ６）。

【００５２】

次に、索引生成器２４は、キーワード対応テーブル５２のエントリキーワードとその読み（又はスペル）に基づいて、文字列グラフを生成し、それを縮退し、その結果を索引情報テーブル６１に格納する（ステップＳ７）。縮退については後述する。

【００５３】

次に、索引生成器２４は、索引情報テーブル、関連文書テーブル、関連キーワードテーブル及び文書情報テーブルに基づいて索引を作成する（ステップＳ８）。索引は例えばハイパーテキストで作成される。作成された索引は、表示装置１３に表示されてもよい。

【００５４】

生成された索引は索引アクセス部２５を介して表示装置１３に出力され、利用者は表示された索引を用いて入力をする。索引アクセス部２５は利用者のアクセス状況を示す情報をアクセス解析器２６に出力する。アクセス解析器２６はアクセス状況を示す情報を解析してアクセスログ３４を作成する（不図示）。

【００５５】

以下、文書検索装置のリンク重要度付与器がリンク重要度を算出する手順について説明する。

まず、本実施形態において、リンク重要度付与器２１はリンク重要度を付与する際に、その文書データのリンク関係、ＵＲＬ及びキーワードを利用する。なお、リンク関係に基づいて判定される文書データの重要度をリンク重要度という。リンク重要度を判定する際の基本的な考え方は以下の通りである。

１．類似度の低いＵＲＬから多くリンクされている文書データ（ページ）は重要である。

【００５６】

例えば、一般に、同一サイト内に設けられた複数のＷＷＷページ（ページ）はそのサイト内の他のページにリンクされているが、それらのページのＵＲＬは相互に類似する。従って、類似度の高いＵＲＬからリンクされているページの重要度は低いと推定できるからである。

２．多くのページからリンクされているページほど重要なページであり、重要なページからリンクされている、ＵＲＬの類似度の低いページは重要である。

【００５７】

例えば、有名なディレクトリサービス等及び官公庁等は多くのページからリンクされているが、このような重要なページからリンクされている文書の方が、個人が開設するページやそのコンテンツのエントリページからリンクされている文書よりも重要度が高いと考えられるからである。また、多くのページやミラーサイトを抱えるサービス（サイト）に設けられたページ等はそのサイト内でリンクされていることが多い。そのため、従来技術において、同じサイトのページが多く検索されてしまうという問題があった。しかし、そのサイト内のページのＵＲＬは例えばドメインが同じ等大抵類似しているため、「ＵＲＬの類似度の低いページは重要である」という考え方を導入すれば、この問題を解消することが可能となる。

３．ＵＲＬの類似度は、サーバアドレス、パス、ファイル名の全てが異なるものが最も小さく、ミラーサイトや同一サーバ内のページは類似度が高くなるように、ＵＲＬの字面情報から定義する。

【００５８】

上述の３つの考え方を導入することにより、全てのリンク関係を同等に扱わな

いでリンク重要度に応じた重みをリンク関係に与えることとしている。より具体的には、リンクの重みをリンク元とリンク先文書のＵＲＬの類似度の逆数として与えることとしている。これにより、単純にリンクされている数だけで文書の重要度を判定しすべてのリンクを同等に扱っている従来技術における問題、つまり、大量のページやミラーサイトを抱えるサーバ（サイト）や個人が単に量が多いというだけで重要度が高いことになるという問題を解決することが可能となる。また、故意にサイト内のページを増やしてリンクを設定することによりページの重要度を上げようとしても、同じサイトのページのＵＲＬ類似度は高いために、ページの重要度を上げることは従来よりも困難になるという効果もある。

【００５９】

以下、リンク重要度付与器２１におけるリンク重要度の算出についてより詳しく説明する。

リンク重要度の算出対象となるページ集合をＤＯＣ＝｛p_1, p_2,....p_N｝、

ページｐのリンク重要度をW_p、

ページｐのリンク先のページ集合をＲef(p)、

ページｐのリンク元のページ集合をＲefed(p)、

ページｐとｑのＵＲＬ類似度をsim(p,q)、

相異度をdiff(p,q) ＝1/sim(p,q)とすると、

ページｐからｑにリンクが張られているとした時、そのリンクの重みlw(p,q)を以下の（１）式で定義する。

【００６０】

【数３】

$$lw(p,q) = diff(p,q) / \sum_{i \in Ref(p)} diff(p,i)$$

$$= \frac{1}{sim(p,q) \sum_{i \in Ref(p)} \frac{1}{sim(p,i)}} \qquad \cdots\cdots\cdots(1)$$

【００６１】

この（１）式から分かるように、lw(p,q) は、ｐとｑのＵＲＬの類似度sim(p,q)が低いほど、また、ｐからのリンク数がより少ないほど大きくなる。

各ページのリンク重要度は、各ｐ∈ＤＯＣに対して、C_q を定数（重要度の下限であり、ページによって異なる値を与えてもよい。）として、

【００６２】

【数４】

$$Wq = Cq + \sum_{p \in Refed(q)} W_p * lw(p,q) \qquad \cdots\cdots\cdots(2)$$

【００６３】

という連立一次方程式の解として定義する。リンク重要度付与器２１は、この連立一次方程式を解くことにより、リンク重要度を各ページに付与する。なお、このような連立一次方程式の解法については、既存のアルゴリズムが多数存在するため、説明は省略する。（１）式中のＵＲＬ類似度sim(p,q)はリンク重要度付与器２１に備えられたＵＲＬ類似度計算器２７により、計算される（後述）。（１）式及び（２）式から、上述の考え方が実現されていることを読み取ることができる。すなわち、（１）式から類似度が低ければ、重みｌｗは大となるから（２）式から類似度の低いＵＲＬから多くリンクされている文書ページは重要となる。また、（２）式から多くのページからリンクされているページほど重要なページとなる。

【００６４】

さらに、（２）式から重要なページ（Ｗｑ）からリンクされているＵＲＬの類似度の低いページ（リンクの重みｌｗの高い）は重要である。以下、図７及び図８を用いて、（１）式及び（２）式の示す考え方についてより詳しく説明する。

【００６５】

図７は、（１）式及び（２）式の示す内容を模式的に示す。図７において、まるは各ページ、矢印はリンク関係、矢印の方向はリンクの方向、矢印の太さがリ

ンクの重みを示す。図７に示すように、ページp_1、p_2及びp_3からページｑへリンクがはられている。ページp_1はページｑ以外の２つのページr_1及びr_2にもリンクし、ページp_3は他の２つのページs_1及びs_2からリンクされている。

【００６６】

ここで、各ページのＵＲＬ類似度は、

sim(p1,q) =sim(p1,r1)=sim(p2,r1)= 1

sim(p2,q) = 2、（つまり、ページp_2及びｑのＵＲＬは若干類似する）

sim(p3,q) = 1、sim(s1,p3)=sim(s2,p3)=3

（つまり、ページs_1、s_2及びp_3のＵＲＬは類似する）

とする。

【００６７】

（１）式及び（２）式を図７のような場合に適用すると、

各ページp_1、p_2、p_3、s_1及びs_2のリンクの重みは、以下のようになる。

【００６８】

lw(p1,q)＝１／｛１×（１＋１＋１）｝＝１／３

lw(p2,q)＝１／｛２×（１／２）｝＝１

lw(p3,q)＝１

lw(s1,p3) ＝lw(s2,p3) ＝１／３

故に、（１）式及び上述の計算からリンク先の多いページp_1のリンクの重みlw(p1,q)は小さくなることが分かる。また、同様に（１）式及び上述の計算からＵＲＬ類似度が大きいほどリンクの重みは小さくなることも分かる。

【００６９】

また、ページｑのリンク重要度W_qは以下のようになる。

$$W_q = C_q + \{lw(p1,q)\times W_{p1} + lw(p2,q)\times W_{p2} + lw(p3,q)\times W_{p3}\}$$

$$= C_q + \{(W_{p1}/3) + W_{p2} + W_{p3}\}$$

$$W_{p1} = C_{p1}$$

$$W_{p2} = C_{p2}$$

$W_{p3}=C_{p3}+\{lw(s1,p3)\times W_{s1}+lw(s2,p3)\times W_{s2}\}$

$=C_{p3}+(W_{s1}+W_{s2})/3$

故に、ページp_1及びp_2と比べてより多くのページからリンクされているページp_3のリンク重要度W_{p3}は大きくなっている。また、ページqのリンク重要度W_qをみれば、重要なページと思われるから、ＵＲＬ類似度の低いページ、すなわちリンク重み$lw(p3,q)=1$、いいかえるとリンク重み（リンク重みの最大値は＝）が大きいページほど、リンク重要度が大きくなることが分かる。また、（２）式及び上述の計算から、ＵＲＬの類似している同じサイト内のページからのリンクの重みを、他のＵＲＬの類似していないページからのリンクの重みよりも軽く扱うことが分かる。これにより、大量のページを抱えるサイトのページが検索結果の中に多く出てくるという問題を解決できていることが分かる。

【００７０】

図８は、（１）式及び（２）式の示す内容を模式的に示す。図８においても、図７と同様に、まるは各ページ、矢印はリンク関係、矢印の方向はリンクの方向、矢印の太さがリンクの重みを示す。さらに、図８において網掛されているまるはＵＲＬ類似度が高いページを示す。図８（ａ）及び８（ｂ）において、ページqはページp_1、p_2及びp_3からリンクされている。さらに、図８（ｂ）において、ページqはページp_1、p_2及びp_3のＵＲＬは類似しており、ＵＲＬ類似度$sim(pi,q)=n+1$である（ｎは整数）。図８（ａ）及び図８（ｂ）それぞれについて（１）式及び（２）式を適用する。図８（ａ）の場合は以下のようになる。

【００７１】

各ページのリンクの重み

$lw(pi,q)=1/sim(pi,q)=1$　（ＵＲＬが非類似）

ページqのリンク重要度W_q

$W_q=C_q+(W_{p1}+W_{p2}+W_{p3})$

図８（ｂ）の場合は以下のようになる。

【００７２】

各ページのリンクの重み

$lw(pi,q) = 1 / sim(pi,q) = 1 / (n+1)$　（ＵＲＬが類似）

ページｑのリンク重要度W_q

$W_q = C_q + (W_{p1} + W_{p2} + W_{p3}) / (n+1)$

故に、図８（ａ）及び８（ｂ）それぞれの計算結果を比較すると、ＵＲＬ類似度sim(p,q)が高い場合には、被リンク数が多くてもページｑのリンク重要度W_qが小さくなることが分かる。延いては、ＵＲＬ類似度を導入することにより、大量のページを抱えるサーバ（サイト）等が単にページの量が多いというだけで重要度が高いことになるという問題を解決していることが分かる。

【００７３】

次に、（１）式及び（２）式中のページｐとｑのＵＲＬ類似度sim(p,q)について説明する。ＵＲＬ類似度は、リンク重要度付与器２１に備えられたＵＲＬ類似度計算器２７により算出される。

【００７４】

一般に、ページのＵＲＬは、サーバアドレス、パス、ファイル名の三種類の情報から構成される。例えば、ＷＷＷページのＵＲＬ、

http://www.flab.fujitsu.co.jp/hypertext/news/1999/product1.html は、サーバアドレス（www.flab.fujitsu.co.jp）、パス（hypertext/news/1999 ）、ファイル名（product1.html）の３種類の情報から構成される。

【００７５】

また、サーバアドレスは、さらに“.”により階層化されており、後ろに行くにしたがって、段々広くなる。例えば、サーバアドレスがwww.flab.fujitsu.co.jpであれば、後ろから、日本(jp)、会社(co)、富士通(fujitsu) 、研究所(flab)、マシン(www) という階層を表している。

【００７６】

本実施形態では、与えられた２つのページｐ及びｑのＵＲＬ類似度を、上記の三種類の組合せにより定義する。類似度sim(p,q)として、例えば、以下に述べるドメイン類似度sim __domain(p,q) 及び融合類似度sim __merge(p,q)が考えられる。

【００７７】

ドメイン類似度sim __domain(p,q) は、ドメインの類似に基づいて算出される。ドメインとは、サーバアドレスの後半部分であり、会社や組織を表す。サーバアドレスが.com、.edu、.org等で終わる米国サーバの場合はサーバアドレスの後ろから２つめまで、サーバアドレスが.jp 、.fr 等で終わる他国のサーバの場合はサーバアドレスの後ろから３つめまでがドメインに相当する。例えば、www.fujitsu.com のドメインはfujitsu.com であり、www.flab.fujitsu.co.jpのドメインはfujitsu.co.jp である。

【００７８】

ページｐとページｑのドメイン類似度は以下の（３）式により定義される。

sim __domain(p,q) = １／α　　（ｐ、ｑが同一ドメインの場合）

= １　　（ｐ、ｑが異なるドメインの場合）

ここで、αは定数で、０より大きく１より小さい実数値を取るとする。図９はインターネットから収集した約３００万ＵＲＬ間のリンク関係にドメイン類似度sim __domain(p,q) の概念を導入してリンク重要度を計算した場合を示す。図９において、横軸が、リンク重要度が高い順から上位何ページかを示す数、縦軸が上位ページの中に含まれる異なるドメインを持つページの数を示し、系列１から５は、それぞれ順にα＝０．１、０．２、０．３、０．５、０．７及び１．０の場合を示す。図９に示すように、リンク重要度が高い上位１０万件のページ中に含まれる異なるドメインを持つページの数は、α＝１の場合（ＵＲＬ類似度を導入しない従来の場合に相当する）は４０００件であり、α＝０．１の場合は５５００件である。従って、αが小さくなるほど、異なるドメインを持つページがリンク重要度の上位に上がってくることが分かる。これは、αが小さいほど類似度sim __domain(p,q) が大きくなり、ＵＲＬ類似度sim __domain(p,q) が大きいほど、リンクの重みlw(p,q) が小さくなり、延いては、リンク重要度W_q が小さくなるため、ＵＲＬ類似度が大きくなるほどページには小さいリンク重要度が付与されるからである。つまり、sim __domain(p,q) の概念を導入することにより、異なるドメインを持つページが検索されやすくなっている、言い換えると、同じドメインを持つページは検索されにくくなっていることが分かる。

【００７９】

sim(p,q)として、前述の三種類の情報を融合した類似度sim __merge(p,q)を次のように定義する。

sim __merge(p,q)＝（サーバアドレスの類似度）＋（パスの類似度）＋（ファイル名の類似度）

以下、右辺の各項の算出方法について説明する。

【００８０】

サーバアドレスの類似度は、アドレスの階層を後ろから見ていき、ｎレベルまで一致した場合、類似度を１＋ｎとする。例えば、www.fujitsu.co.jp とwww.flab.fujitsu.co.jpは３レベルまで一致しているので４となる。www.fujitsu.co.jp とwww.fujitsu.com は１レベルも一致していないので（一致０レベル）、類似度は１である。

【００８１】

パスの類似度は、先頭からパスの”/”で区切られた要素毎に比較し、一致したレベルまでを類似度とする。例えば、/doc/patent/index.htmlと/doc/patent/1999/2/file.htmlとは、２レベルまで一致しているので類似度は２である。

【００８２】

ファイル名の類似度は、ファイル名が一致する場合、類似度１とする。

上記は、以下のような考え方に基づいている。

１．往々にして、同じような文書を同一ディレクトリに入れるため、同一サーバでパスも同じＵＲＬは内容が似ていることが多い。

２．アクセスを分散させるために設けられるミラーサイトは類似度が高い。これらは、サーバアドレス部分だけが異なり、残りのパスやファイル名は同じ場合が多い。

３．サーバアドレス、パス、ファイル名が全てことなるＵＲＬは、類似度が低い。

【００８３】

このsim __merge(p,q)によっても、ＵＲＬが似通ったページが検索されることを防ぐことができる。従って、lw(p,q) の中にsim(p,q)又はdiff(p,q) という概念を導入することにより、大量のページを抱えるサーバ（サイト）や個人が単に

量が多いというだけでリンク重要度が高いことになるという従来の問題を解消することができる。

【００８４】

なお、上述のリンク重要度W_pは、後述の関連度算出にも用いられる。

以下、文書検索装置のキーワード文書関連度計算器２３が関連度を算出する手順について説明する。

【００８５】

まず、キーワードから文書への索引を作るにあたっては、キーワードと文書への関連度が必要である。関連度については、以下のように考える。

１．キーワードを多く含んでいる、つまりキーワードを含む度合いが高い文書ほど関連度が大きい。２．重要度の高い文書ほど関連度が大きい。

３．あるキーワードの関連文書は一定個数以下であることが望ましい。（一つのキーワードで、1,000 個もの関連文書が得られるのはまずい）。

【００８６】

本実施形態ではあるキーワードの関連文書は一定個数以内にするために、上記の考え方に加え、以下の考え方を導入する。

４．利用者のアクセスログ解析による関連度：一定期間内に当該キーワードからよくアクセスされた文書ほど関連度が高い。

５．文書のリンク重要度による関連度：当該キーワードを含む文書のうち、リンク重要度の高いものは関連度が高い。

【００８７】

上述の考え方を導入して、あるキーワードｗに対してページｐの関連度を以下の（３）式で与える。

$$Rel(p,w) = TF(p,w) * \log W_p * \log(AC(p,w) + 2) \quad \cdots (3)$$

ここで、ＴＦ（p,w)はｐにおけるｗの出現個数、

W_p はｐのリンク重要度であり、上述の（２）式のW_p に相当する。

【００８８】

ＡＣ(p,w) は、一定期間（例えば１月間又は１週間）の間にキーワードｗからｐにアクセスされた回数である。

各キーワードに対して、上記Ｒｅｌ(ｐ,ｗ) の値の大きいものから一定個数をそのキーワードの関連ページとする。

【００８９】

キーワードを含む度合いに加えて、ＵＲＬ類似度という概念を導入したリンク重要度Ｗ(ｐ) 及び利用者のアクセスログ解析を関連度の算出に用いている。これにより、ページの関連度を高くするための条件が多くなるため、意図的に第三者があるページの関連度を高くするようにページ内容を操作することはさらに困難になる。

【００９０】

以下、索引生成器２４が生成する索引、つまり、検索のためのキーワードを選択するインターフェースについて説明する。本実施形態の選択インターフェースは、キーワードの読みの断片を順次クリックするだけでキーワードに辿りつけるインタフェースであり、以下の特徴がある。

・一画面には、キーワードの読みの断片（文字または文字列）と、それまで選択したキーワード読みに相当するキーワードの一部が表示される。

・利用者は、画面上のキーワードの読み断片（文字または文字列）を次々クリックすることで、キーワードに辿りつける。

・１画面に表示されるキーワードは、一定個数以内に制限することができる。

【００９１】

前述の従来技術では、常に１「文字」をクリックして読みを選択しているが、本実施形態では必ずしも文字でなく、「文字列」の場合もあり得る。これにより、キーワードに辿りつくまでの無駄なクリックを減らすことが可能である。さらに、１画面に表示されるキーワードは一定個数以内であるため、キーワードを選択することが用意となる。また、１画面に表示されるキーワードは一定個数以内であることは、表示装置の画面の大きさが制限されている携帯電話のような移動端末でのキーワード選択に有用である。これらを実現するために、索引生成器２４において以下のようにしている。

１．キーワードの読み（またはスペル）の統一。場合によっては「ー」（長音）の除去や「ぃ」のような拗撥音を「あ」「い」のような表記に統一する。

２．キーワードとその読み（またはスペル）から、読み文字をノード、キーワード集合をリーフとする有効グラフ（文字列グラフ）を作る。

３．上記グラフで、以下の縮退操作を行う（下を参照）。

【００９２】

(a) リーフへのパスの縮退

(b) 中間パスの除去

(c) 子ノードのキーワードを親ノードにくりこみ、子ノードの削除

４．縮退後のグラフを元に、読み入力インタフェースを作成する。

【００９３】

以下、索引生成器２４においてキーワード文字列グラフ作成以下の手順をどのように行っているか説明する。

キーワード文字列グラフとは、キーワードの読みを表現する有向ラベル付きグラフである。図１０は、キーワード文字列グラフの一例を示す。

【００９４】

キーワード文字列グラフは、

（Ｎ，Ｃ，ＫＷ，ｔ，ｎｋ，ｙｏｍｉ）

の６つ組で表現される。ここで、

Ｎはノードの集合、

Ｃはかな文字の集合、

ＫＷはキーワードの集合、

ｔは$N * C^{+} \rightarrow N$ノード遷移関数。また、C^{+}はラベル、つまり、１以上のかな文字の並び（図９（ａ）等の文字列グラフでは実線矢印で示される）

ｎｋは$N \rightarrow W^{+}$ノードに割り当てられたキーワード（図９（ａ）等では点線で示される）

ｙｏｍｉは$N \rightarrow C^{+}$ノードの「読み」

である。

【００９５】

図１０（ａ）を例に取ると以下のようになる。なお、ｙｏｍｉは自明なので省略）

Ｎ＝｛トップ、「あ」，「あい」，「あいぼ」，「あいぼう」，「あいぼり」，「あお」、「あおぞ」、「あおぞら」｝

Ｃ＝｛’あ’，．．．，’ん’｝

ＫＷ＝｛’青’，’蒼’，’青空’，’アイボリー’｝

ｔ（トップ，あ）＝「あ」，

ｔ（あ，い）＝「あい」，

ｔ（あ，お）＝「あお」，

ｔ（あい，ぼ）＝「あいぼ」，

ｔ（あいぼ，う）＝「あいぼう」，

ｔ（あいぼ，り）＝「あいぼり」，

ｔ（あお，ぞ）＝「あおぞ」，

ｔ（あおぞ，ら）＝「あおぞら」，

ｎｋ（あいぼう）＝｛’相棒’｝

ｎｋ（あいぼり）＝｛’アイボリー’｝

ｎｋ（あお）＝｛’青’’蒼’｝

ｎｋ（あおぞら）＝｛’青空’｝

索引生成器２４はキーワードと読みが与えられた場合、それらに基づいて初期キーワード文字列グラフを生成する。図１１は初期キーワード文字列グラフの生成手順を示す。図１１を用いて索引生成器２４における初期キーワード文字列グラフの生成手順を説明する。なお、図１２は、初期キーワード文字列グラフの生成手順を実現するアルゴリズムの一例を示す。

【００９６】

まず、キーワードの集合ＫＷを作成する（ステップＳ１１）。次に、作成された集合ＫＷ内が空か否か判定する（ステップＳ１２）。集合ＫＷ内が空である場合（ステップＳ１２、ｙｅｓ）、文字列を作成する必要はないため、処理を終了する。集合ＫＷ内が空でない場合（ステップＳ１２、ｎｏ）、次のステップに進む。

【００９７】

次に、集合ＫＷからあるキーワードｕを抜き出す（ステップＳ１３）。そして

、このキーワードｕの読みｙｏｍｉ（ｕ）及び、この読みｙｏｍｉ（ｕ）のノードであるｎｋ｛ｙｏｍｉ（ｎ）｝を設定し、このノードｎｋ｛ｙｏｍｉ（ｎ）｝を末端ノードとして追加する（ステップＳ１４）。

【００９８】

ステップＳ１４の処理をキーワードｕの文字列長分繰り返したか否か、つまり、キーワードｕが空になったか否か判定する（ステップＳ１５）。キーワードｕが空になったか場合（ステップＳ１５、ｙｅｓ）、キーワードｕについては必要な処理は終了したため、ステップＳ１２へ戻り集合ＫＷから別のキーワードをとってステップＳ１３以下の手順を繰り返す。キーワードｕが空になっていない場合（ステップＳ１５、ｎｏ）、キーワードｕの末尾文字を切り取る（ステップＳ１６）。続いて、ノードを１つ前の親ノードに遷移して追加する（ステップＳ１７）。さらにキーワードｕの切り取られた末尾文字の前の文字に注目し（ステップＳ１８）、ステップＳ１５に戻る。

【００９９】

この手順により、集合ｎｋとしてノードに割り当てられたキーワードリストが得られ、ｔとしてあるノードの下位に位置する下位ノードのリストが得られる。

図１６（ａ）は上述の手続により作成された初期キーワード文字列グラフを示す。図１６（ａ）は、

蒼：あお、

青：あお、

青空：あおぞら、

相棒：あいぼう、

アイボリー：あいぼり

というキーワードと読みデータから生成した初期キーワード文字列グラフである。また、同様に、図９（ａ）は図１１に示すinit__kw__graph() アルゴリズムにおいて

@KW ＝｛蒼、青、青空、相棒、アイボリー｝

yomi｛蒼｝＝あお、yomi｛青｝＝あお、yomi｛青空｝＝あおぞら、yomi｛相棒｝＝あいぼう、yomi｛アイボリー｝＝あいぼり、

としたものである。

【０１００】

索引生成器２４は、初期キーワード文字列グラフを作成すると、次に、文字列の縮退を行う。以下、文字列の縮退について説明する。縮退は、

１．中間ノードの縮退

２．末端ノードのくりこみ

の２種類の操作からなる。

【０１０１】

まず、索引生成器２４における中間ノードの縮退処理について説明する。図１３は中間ノードの縮退処理の手順を示す。以下、図１３を用いて中間ノードの縮退処理の手順について説明する。なお、図１４は、中間ノードの縮退処理を実現するアルゴリズムの一例を示す。

【０１０２】

まず、ノードの集合Ｎを作成する（ステップＳ２１）。続いて、集合Ｎ内が空か否か判定する（ステップＳ２２）。集合Ｎ内が空である場合（ステップＳ２２、ｙｅｓ）、ノードを縮退させる必要はないため、処理を終了する。集合Ｎ内が空でない場合（ステップＳ２２、ｎｏ）、集合Ｎ内のノードｎを取得する（ステップＳ２３）。取得したノードｎについて、そのノードｎに続く後継ノードが１つしかなく、かつノードｎ内にキーワードもないという２つの条件を満たすか否か判定する（ステップＳ２４）。２つの条件を満たす場合（ステップＳ２４、ｙｅｓ）、ノードｎは縮退可能であるため、ノードｎをキーワード文字列グラフから削除し、ステップＳ２２に戻る（ステップＳ２５）。２つの条件の両方を満たさない場合（ステップＳ２４、ｎｏ）、ノードｎは縮退できないので、ノードｎを削除せずにステップＳ２２に戻る。

【０１０３】

上述のようにしてキーワード文字列グラフにおいて、「ノードに割り当てられたキーワードがない」又は、「後継ノード（子ノード）が１つのみ」であるという２つの条件を満たす中間ノードは縮退させることとしている。図１０（ａ）の初期キーワード文字列グラフでは、ノード「あい」及びノード「あおぞ」が「ノ

ードに割り当てられたキーワードがない」又は、「後継ノード（子ノード）が１つのみ」であるという２つの条件を満たす中間ノードに相当する。図１０（ｂ）は、図１０（ａ）に示す初期キーワード文字列グラフの中間ノードを縮退させた結果を示す。図１０（ｂ）において、中間ノード「あい」及び「あおぞ」は削除されている。また、同様に、図１４に示すproc__shrink__middle()アルゴリズムにおいて、

ｔ｛''｝＝あ＋

ｔ｛あ｝＝あいぼ＋あお＋

ｔ｛あいぼ｝＝あいぼう＋あいぼり＋

ｔ｛あお｝＝あおぞ＋

ｎｋ｛あいぼう｝＝相棒＋

ｎｋ｛あいぼり｝＝アイボリ＋

ｎｋ｛あお｝＝青＋蒼＋

ｎｋ｛あおぞら｝＝青空＋

となる。

【０１０４】

次に、索引生成器２４における末端ノードの縮退処理について説明する。図１５は末端ノードの縮退処理の手順を示す。以下、図１５を用いて末端ノードの縮退処理の手順について説明する。なお、図１６は、末端ノードの縮退処理を実現するアルゴリズムの一例を示す。

【０１０５】

まず、全てのノードの集合Ｎを作成する（ステップＳ３１）。次に、ノードをそのノードに属するキーワード数の順にソートする（ステップＳ３２）。整数ｉを１に設定する（ステップＳ３３）。次に、整数ｉが集合Ｎ内のノードの数よりも少ないか否か判定する（ステップＳ３４）。整数ｉが集合Ｎ内のノードの数よりも少ないのではない場合（ステップＳ３４、ｎｏ）、さらに、末端ノードの縮退があるか否か判定する（ステップＳ３５）。末端ノードの縮退がない場合（ステップＳ３５、ｎｏ）、処理を終了する。末端ノードの縮退がある場合（ステップｓ３５、ｙｅｓ）、ステップＳ３３に戻る。

【０１０６】

整数ｉが集合Ｎ内のノードの数よりも少ない場合（ステップＳ３４、ｙｅｓ）、集合Ｎのｉ番目のノードｎを取得する（ステップＳ３６）。取得されたノードｎが末端ノードであるか否か判定する（ステップＳ３７）。取得されたノードｎが末端ノードである場合（ステップＳ３７、ｙｅｓ）、ステップＳ３８に進む。取得されたノードｎが末端ノードでない場合（ステップＳ３７、ｎｏ）、ノードｎは縮退すべきノードではないため、整数ｉを１つインクリメントし（ステップＳ４１）、ステップＳ３４に戻る。

【０１０７】

次に、ノードｎの親ノードｐを取得する（ステップＳ３８）。次に、親ノードｐに属するキーワードの数と子ノードｎに属するキーワードの数の和が予め与えられたキーワードの数の上限を超えるか否か判定する（ステップＳ３９）。

【０１０８】

親ノードｐに属するキーワードの数と子ノードｎに属するキーワードの数の和が上限値を超えない場合（ステップＳ３９、ｙｅｓ）、子ノードｎを削除し（縮退させる）、ノードｎに属するキーワードを親ノードｐに属するキーワードとする（ステップＳ４０）。続いて、整数ｉを１つインクリメントして（ステップＳ４１）、ステップＳ３４へ戻る。

【０１０９】

親ノードｐに属するキーワードの数と子ノードｎに属するキーワードの数の和が上限値を超える場合、子ノードｎを縮退させると親ノードのキーワード数が多すぎることになるため、子ノードｎを縮退させずに、ステップＳ４１に進む。

【０１１０】

このようにして末端ノードにあるキーワード情報を、一つ上の親ノードに移すことにより、木（ノードの連鎖）の深さを減らし、利用者が少ないクリックでキーワードに到達できるようになる。しかし、あまり沢山親ノードに子ノードのキーワードを移しすぎては、一つのノードに大量のキーワードが割り当てられることになるので、利用者がその中からキーワードを選択することが面倒になってしまう。この問題を回避するため、パラメータwords __max を与え、１ノードに属

するキーワードがそのパラメータwords __max より少なくなるようにする。

【０１１１】

図１０（ｃ）は、図１０（ｂ）のキーワード文字列グラフをパラメータwords __max ＝４の場合に末端ノード縮退処理した結果を示す。図１０（ｂ）において、末端ノード「あいぼう」及び「あいぼり」はそれぞれ１つのキーワードしか持たない。末端ノード「あいぼう」及び「あいぼり」の親ノード「あいぼ」は２つの子ノードを持ち、キーワードはない。従って、親ノード「あいぼ」と子ノード「あいぼう」及び「あいぼり」のキーワードの和はwords __max ＝４より少ない。故に、子ノード「あいぼう」及び「あいぼり」は縮退可能であるため、図１０（ｃ）において子ノード「あいぼう」及び「あいぼり」は削除され、これら子ノードのキーワードは親ノード「あいぼ」に移されている。また、同様に、図１６に示すproc__shrink__leaf()アルゴリズムにおいて、

ｔ｛''｝＝あ＋

ｔ｛あ｝＝あいぼ＋あお＋

ｔ｛あお｝＝あおぞ＋

ｎｋ｛あいぼ｝＝相棒＋アイボリー＋

ｎｋ｛あお｝＝青＋蒼＋

ｎｋ｛あおぞら｝＝青空＋

となる。

【０１１２】

図１７は、末端ノード縮退処理をしたキーワード文字列グラフの更なる一例を示す。図１７において、親ノード「かいせ」には３つの末端ノード「かいせい」、「かいせき」及び「かいせつ」がある。末端ノード「かいせい」及び「かいせき」に属するキーワードはそれぞれ「快晴」及び「解析」であるため、末端ノード「かいせい」及び「かいせき」は縮退可能である。また、末端ノード「かいせつ」に属するキーワードは「解説」及び「開設」の２つであるため、末端ノード「かいせつ」も縮退可能である。従って、図１７の場合は２通りの縮退の方法が考えられる。しかし、前者の方が総ノード数を減らすことができる。本実施形態において、末端ノードをそのノードに属するキーワード数の順にソートしている

。これにより、前者のような効率的な縮退をすることを可能にしている。

【０１１３】

次に、索引生成器が作成する索引の例について図１８から図２５を用いて説明する。図１８は索引トップ画面から索引中間画面及びキーワード情報画面を経て分所ページに至るまでの索引画面の遷移を示す。図１８を用いて表示装置に表示される索引画面の遷移について説明する。図１８において、まず、索引トップ画面が表示される。利用者が索引トップ画面からキーワードの読み又はスペルの最初の部分を選択すると、索引中間画面に遷移する。利用者が索引中間画面からキーワードの読み又はスペルの次の部分を選択すると、さらなる索引中間画面に遷移する。この選択を繰り返した結果、利用者が検索を希望するキーワードに辿りつき、そのキーワードを選択すると、キーワード情報画面に遷移する。利用者がその他のキーワードを選択すると、さらなるキーワード情報画面に遷移する。利用者が閲覧を希望する文書ページのタイトルを選択すると、その文書ページへリンクし、その文書ページに遷移する。なお、利用者の選択操作は、マウスによるクリックやペン等のポインティングにより行うようにすることも可能である。各画面は、例えばハイパーテキストで作成されてもよい。

【０１１４】

図１９は索引トップ画面の一例を示す。索引トップ画面には、図４に示す索引情報テーブル６１のトップ以下の文字（列）が表示される。図１９において、５０音、アルファベット及び０から９までの数字が表示されている。利用者は検索したいキーワードの読みまたはスペルの最初をクリックすることで先の画面に進むことができる。

【０１１５】

図２０は索引トップ画面のさらなる例を示す。図２０において、キーワードの読み付けをする際に表記統一した又は・及びノードの縮退をした結果、例えば、５０音の「だ」行の「ぢ」及び「づ」が索引から除かれていることが分かる。同様に、アルファベットの「Ｙ」及び「Ｚ」が除かれていること等が分かる。

【０１１６】

図２１は索引中間画面の一例を示す。図２１は、索引トップ画面において「あ

」を選択した場合の索引中間画面である。上半分は「あ」の後に継続する文字列、下半分にはその他のキーワードが表示されている。索引中間画面は、図４の索引情報テーブル６１及び図３のキーワード対応テーブル５１（キーワードから、キーワードＩＤを得る）から生成できる。

【０１１７】

図２１において、「あ」に継続する文字列として「いぼ」、「え」「お」などがある。例えば、「いぼ」を選択すると「あいぼ」に移る。その他のキーワードとして、「愛」「愛犬」などが一定個数（例えば20以内）表示されている「あ」から始まっていて、継続する読みが選べないキーワードは全てここに表示されている。逆にキーワードも表示されず、継続する読みもない場合には、そのキーワードは索引に含まれていないことがわかる。

【０１１８】

図２２は、索引中間画面の更なる例を示す。図２２は、索引トップ画面において「い」を選択した場合の索引中間画面である。上半分は「い」の後に継続する文字列、下半分にはその他のキーワードが表示されている。図２２において、「あ」に継続する文字列として「いぼ、「えろ」などがある。その他のキーワードとして、「イオン」「イネーブル」などが一定個数（例えば20以内）表示されている。

【０１１９】

図２３は、索引中間画面のさらなる例を示す。図２３において、「いべんと」が選択された場合の索引中間画面が示されている。ノード「イベント」の子ノードはないため、「いべんと」の後に継続する文字列は表示されず、キーワードが表示されている。利用者は、この画面から入力したいキーワードを選択する。この画面に含まれていない「いべんと」を含むキーワードは索引に含まれていないことが分かる。

【０１２０】

従来技術において読みは１文字ずつしか選択できないため、互い文字列を入力するためには何度も選択操作を繰り返す必要があった。しかし、本実施形態によれば、ノードを縮退させるため、継続する文字列を必ずしも１文字ずつ選択する

必要はない。例えば図２０の文字列「いぼ」又は図２１の文字列「えろ」のように、２文字を一度に選択することも可能である。従って、利用者が選択する回数を低減することができる。

【０１２１】

また、さらに、継続する読みが選べないキーワードは全てここに表示されている。逆にキーワードも表示されず、継続する読みもない場合には、そのキーワードは索引に含まれていないため、利用者が１文字づつ選択していって最後の１文字をするときになって初めて入力しようとしていたキーワードが索引に含まれていないことが分かるといった問題は生じないことになる。

【０１２２】

また、さらに、末端ノードの縮退の際に、パラメータword#maxとして設定された値以上のキーワードは表示されないため、利用者はキーワードの選択の際に表示されたキーワードから選択したいキーワードを探し出すことが比較的楽におかなうことができる。また、表示されるキーワードが一定個数以内にかぎられていることは、表示画面の大きさが限られた携帯電話のような移動端末におけるキーワード選択において有用である。

【０１２３】

さらにまた、検索のインタフェースとして、決まったキーワードをセットに、なるべく少ない労力で辿り着かせることを可能にし、従来技術におけるかな漢字変換とは次のような点が異なる。

・変換キー操作は不要

・変換したいキーワードの読みの全てを入れなくても、そのキーワードを特定できる最低限の情報さえ与えられれば良い。

【０１２４】

従って、例えば、キーワードセットに、「なれ」から始まる語として「ナレッジマネージメント」しか入っていない場合、利用者が「な」「れ」と指定したら、それだけで「ナレッジマネージメント」を表示する。

【０１２５】

図２４は、キーワード情報画面の一例を示す。中間画面でキーワード「圧縮」

をクリックした時のキーワード情報画面が示されている。図２４において、まず、代表語と同義語（「圧縮」「Compress」）とが表示される。これは、図３のキーワードテーブル５１とキーワード対応テーブル５２の情報から得ることができる。画面右上に「あ」とあるのは、辿ってきたパスを示す。これにより、前に戻ったり、ハイパーテキストで迷子になるのを防ぐことができる。また、画面には各文書のタイトル、リンク情報、他のキーワードが表示されている。優先度の高い順に一定個数（例えば20文書）以内が表示されるので、利用者はこの中から文書を選択するのが困難ということはない。文書ＩＤのリストは、図４の関連文書テーブル６２から得ることができる。各文書ＩＤについての情報は、図２の文書情報テーブル４１に入っている。他のキーワードは、図４の関連キーワードテーブル６３から得る。利用者は閲覧を希望する文書情報を選択すると、その文書へリンクして文書を画面に表示させることができる。

【０１２６】

図２５は、キーワード情報画面のさらなる例を示す。中間画面でキーワード「イベントカレンダー」をクリックした時のキーワード情報画面が示されている。画面右上に「トップ」－「イ」－「イベント」とあるのは、辿ってきたパスを示す。それぞれのパスを選択することにより、利用者は先に見た画面に戻ることができる。

【０１２７】

図２６は第２実施形態に係わるイントラネット文書検索装置の構成を示す。図２６において、図１に示す第１実施形態の構成に加えて、収集装置８１及び同義語辞書８２を更に備える。収集装置８１はイントラネット（又はインターネット）から大量の文書を収集する、例えばWebロボットである。同義語辞書（同義語データ）８２は同キーワード対応テーブルの情報の一部を格納する。なお、入力装置及び出力装置は、例えば、ＷＷＷブラウザ８３であってもよい。

【０１２８】

収集装置８１はネットワークから文書を自動で収集し、処理装置１１のキーワード抽出器２２は同義語辞書８２を用いて、収集されたページからキーワードを抽出し、ページ内のキーワード出現頻度を集計する。つまり、同義語辞書８２を

用いて重要度の高い文書を自動選別する。これにより、イントラネット（又はインターネット）の大量の文書を自動選別することを可能にする。

【０１２９】

図２７は第３実施形態に係わる特定タイプ文書のイントラネット文書検索装置の構成を示す。図２７において、図２６に示す第２実施形態の構成に加えて、処理装置１１内に文書タイプ判別器９１を備える。文書タイプ判別器９１はイントラネット（及び・又はインターネット）から収集された文書データのリンク関係及びＵＲＬに基づいて、その文書データの文書タイプを判別する。より具体的には、文書タイプ判別器９１はリンク重要度付与器２１内で算出されたＵＲＬ類似度とリンク重要度付与器２１により抽出されたリンク関係が示すリンク／被リンク数に基づいて、（その内容を理解せずに）文書データのコンテンツのタイプを判別する。文書タイプ判別器９１における文書タイプの判別は以下のように行う。

１．ＵＲＬ類似度が一定以上の文書データへのリンクを一定数以上持つページはリンク集である。

２．ＵＲＬ類似度が一定以下の文書データへのリンクを一定数以上持つページはメニュー（エントリ）ページである。

３．ＵＲＬ類似度が一定以下の文書データから一定数以上参照されているページは、メニュー（エントリ）ページである。

４．それ以外で、ＵＲＬ類似度が一定以上文書データへのリンクが一定数以下のページはコンテンツページである。

【０１３０】

このように判別することにより、文書タイプ判別器９１は、かなりの確率で文書データ（WWWページ）の文書タイプ（例えば、メニューページ、リンク集、コンテンツページ）を分けることができる。

【０１３１】

文書タイプ判別器９１は、文書タイプを判別し、判別された文書タイプ９２をキーワード文書関連度計算器２３に出力する。キーワード文書関連度計算器２３は、判別された文書タイプ９２に基づいて特定タイプの文書データを選択し、選

択された文書データについてリンク重要度、ページキーワード及びアクセスログに基づいて文書関連度を計算する。例えば、キーワード文書関連度計算器２３は文書タイプがコンテンツページであると判別された文書データを選択して、これらのコンテンツページについての関連度を計算するようにしてもよい。

【０１３２】

このように、図２６のイントラネット文書検索装置は、索引に掲載するページをここで得た文書タイプ判別器の判別したタイプのいずれかに限定することで、より質の高い文書整理を行うことが可能となる。

【０１３３】

図２８は第４実施形態に係わるリンク集生成システムの構成を示す。図２８において、リンク集生成システムは、収集装置１０１、処理装置１０２及び入出力装置１０７を備える。収集装置１０１はインターネット（又は・及びイントラネット）から収集装置８１は大量の文書データを収集する、例えばWebロボットである。処理装置１０２は、リンク重要度付与器２１、ＵＲＬ文字列判別器１０３、索引生成器２４及びＷＷＷサーバ１０６を備える。リンク重要度付与器２１はＵＲＬの類似度とリンク関係を元に、文書のリンク重要度を算出し、算出されたリンク重要度３１を索引生成器２４へ出力する。

【０１３４】

ＵＲＬ文字列判別器１０３はＵＲＬの文字列上の特徴に基づいて、収集された文書データのコンテンツを（その内容を理解せずに）判別する。ＵＲＬ文字列判別器１０３は、ＵＲＬの文字列上の特徴から例えば、以下のように文書データのコンテンツを判別する。

１．ＵＲＬの文字列にY2K，y2k，year2000を含むような文書データは2000年問題の関連ページである。

２．ＵＲＬの文字列にnews，release，pressを含み、その後で数字列（往々にして日時の情報）を持つ文書データはニュース（プレス）リリースのページである。

３．ＵＲＬの文字列にjava，JAVAを含むページはJava関連ページである。

４．ＵＲＬの文字列にdownload，dwnload，dwnldを含むページは、ソフトウェア

ダウンロード関係のページである。

５．ＵＲＬの文字列にLINUX，linux，Linux を含むページは、Linux 関連ページである。

【０１３５】

このように判別することにより、ＵＲＬ文字列判別器１０３は特定のＵＲＬを判別し、判別された特定ＵＲＬ集合１０４を索引生成器２４に出力する。索引生成器２４は、リンク重要度３１に基づいて、特定ＵＲＬ集合１０４をリンク重要度の大きい順に並べて、上位から一定個数を取り出し、上位ＵＲＬからリンク集を作成し、リンク集１０５としてWWWサーバ１０６に出力する。なお、文字列判別器１０３で得られたＵＲＬが少数の場合には、リンク関係を調べてそれらのリンク元がよく参照（リンク）する他のページを加えて数を増やしてもよい。これは、しばしばWWWで似たようなページは、似たようなリンク集から同時に参照されるためである。WWWサーバ１０６はリンク集を利用者に提供し、利用者はWWWブラウザ１０７を介してリンク集を表示させ、入力指示を行う。

【０１３６】

このようにして、ＵＲＬ文字列に基づいて、文書ページのコンテンツをみないでコンテンツを判別し、この判別結果を用いてリンク集を作成する。これにより、コンテンツの内容にそった質の高いリンク集を簡便に作成することが可能となる。

【０１３７】

図２９は第５実施形態に係わるリンク集生成システムの構成を示す。図２９のリンク集生成システムは、図２８のリンク集生成システムに文書タイプ判別器１１１を更に備える構成となっている。文書タイプ判別器１１１の機能及び動作は、図２７を用いて説明した第３実施形態に係わる文書検索装置の文書タイプ判別器９１と同様である。

【０１３８】

収集装置１０１はインターネット（又は・及びイントラネット）から収集装置８１は大量の文書データを収集し、リンク重要度付与器２１はＵＲＬの類似度とリンク関係を元に文書のリンク重要度を算出し、リンク重要度３１を索引生成期

２４へ出力する。ＵＲＬ文字列判別器１０３はＵＲＬの文字列上の特徴に基づいて、特定のＵＲＬを判別し、判別された特定ＵＲＬ集合１０４を索引生成器２４へ出力する。文書タイプ判別器１１１は、ＵＲＬ類似度とリンク／被リンク数に基づいて、（コンテンツを見ずに）文書データの文書タイプを判別し、判別された文書タイプ１１２を索引生成器２４に出力する。

【０１３９】

索引生成器２４は、文書タイプ１１２に基づいて特定ＵＲＬ集合１０４の中から特定の文書タイプの文書データを選び出す。続いて、選ばれた文書データのリンク重要度３１に基づいて、選ばれた文書データをリンク重要度順の大きい順に並べて、上位から一定個数を取り出し、上位ＵＲＬからリンク集を作成し、リンク集１０５としてWWWサーバ１０６に出力する。WWWサーバ１０６はリンク集を利用者に提供し、利用者はWWWブラウザ１０７を介してリンク集を表示させ、入力指示を行う。

【０１４０】

これにより、コンテンツの内容にそった質の高いリンク集を簡便に作成することが可能となる。

ところで、図１、２６及び２７の文書検索装置並びに図２８及び２９のリンク集生成システムは、図３０に示すような情報処理装置（コンピュータ）を用いて構成することができる。図３０の情報処理装置は、ＣＰＵ１２１、メモリ１２２、入力装置１２３、出力装置１２４、外部記憶装置１２５、媒体駆動装置１２６、及びネットワーク接続装置１２７を備え、それらはバス１２８により互いに接続されている。

【０１４１】

メモリ１２２は、例えば、ＲＯＭ（Ｒｅａｄ　Ｏｎｌｙ　Ｍｅｍｏｒｙ）、ＲＡＭ（Ｒａｎｄｏｍ　Ａｃｃｅｓｓ　Ｍｅｍｏｒｙ）等を含み、処理に用いられるプログラムとデータを格納する。ＣＰＵ１２１は、メモリ１２２を利用してプログラムを実行することにより、必要な処理を行う。

【０１４２】

図１、２６及び２７の文書検索装置並びに図２８及び２９のリンク集生成シス

テムの各処理装置を構成する各機器及び各部は、それぞれメモリ１２２の特定のプログラムコードセグメントにプログラムとして格納される。

【０１４３】

入力装置１２３は、例えば、キーボード、ポインティングデバイス、タッチパネル等であり、ユーザからの指示や情報の入力に用いられる。出力装置１２４は、例えば、ディスプレイやプリンタ等であり、利用者への問い合わせ、処理結果等の出力に用いられる。

【０１４４】

外部記憶装置１２５は、例えば、磁気ディスク装置、光ディスク装置、光磁気ディスク装置等である。この外部記憶装置１２５に上述のプログラムとデータを保存しておき、必要に応じて、それらをメモリ１２２にロードして使用することもできる。

【０１４５】

媒体駆動装置１２６は、可搬記録媒体１２９を駆動し、その記録内容にアクセスする。可搬記録媒体１２９としては、メモリカード、フロッピーディスク、ＣＤ－ＲＯＭ（Ｃｏｍｐａｃｔ　Ｄｉｓｋ　Ｒｅａｄ　Ｏｎｌｙ　Ｍｅｍｏｒｙ）、光ディスク、光時期ディスク等、任意のコンピュータ読み取り可能な記録媒体が用いられる。この可搬記録媒体１２９に上述のプログラムとデータを格納しておき、必要に応じて、それらをメモリ１２２にロードして使用することもできる。

【０１４６】

ネットワーク接続装置１２７は、ＬＡＮ（Ｌｏｃａｌ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）、ＷＡＮ（Ｗｉｄｅ　Ａｒｅａ　Ｎｅｔｗｏｒｋ）等の任意のネットワーク（回線）を介して外部の装置を通信し、通信に伴なうデータ変換を行う。また、必要に応じて、上述のプログラムとデータを外部の装置から受け取り、それらをメモリ１２２にロードして使用することもできる。

【０１４７】

図３１は、図３０の情報処理装置にプログラムとデータを供給することのできるコンピュータ読み取り可能な記録媒体を示している。可搬記録媒体１２９や外

部のデータベース１３０保存されたプログラムとデータは、メモリ１２２にロードされる。そして、ＣＰＵ１２１は、そのデータを用いてそのプログラムを実行し、必要な処理を行う。

【０１４８】

【発明の効果】

本発明によれば、WWWページの重要度を計算する時に、リンク関係だけでなく、ＵＲＬの類似度を考慮に入れることで、多くのページを抱える特定サイトやそのミラーサイトの重要度が過大に評価されにくくして、重要度を計算することが可能となる。これにより、より精度高く重要文書を選定することが可能となる。

【０１４９】

また、本発明により計算される重要度には悪意を持った個人が意図的に操作しにくい性質がある。

また、本発明によれば、索引として、キーワードの読み（又はスペル）の１文字以上の断片を次々クリックするだけで、効率よくキーワード又は文書データにアクセスすることが可能となる。

【０１５０】

また、キーワード索引において、１画面上のキーワードや文書数を一定個数内に収めることができ、利用者はキーワードを選択しやすいだけでなく、携帯端末のような表示領域が限られたメディアでの利用にも有効である。

【０１５１】

また、本発明によれば、文書データ中のキーワード出現頻度と上述の文書データの重要度に基づいて文書関連度を計算し、この関連度の順に文書データへアクセスするリンクを並べることにより、特定のキーワードに関連する良質の文書への迅速なアクセスを可能とするリンク集を作成することが可能となる。

【０１５２】

本発明によれば、ＵＲＬの類似と参照／被参照数を利用して、文書のタイプ（メニュー、リンク集、コンテンツ）を判別することが可能となる。さらに、文書タイプの判別結果に基づいて文書データを選定し、選定された文書データについ

てリンク重要度の算出結果及び・又はキーワード出現頻度を組み合わせて、より質の高い文書へのアクセスを可能とするリンク集を作成することが可能となる。

【０１５３】

また、本発明によれば、特定のＵＲＬを判別することにより、高い精度で特定の分野の文書データを自動的に選定することが可能となる。さらに上述のリンク重要度と判別された特定のＵＲＬに基づいて、特定分野の文書データにアクセスすることを可能とするリンク集が精度良く簡単に作成することが可能となる。

【０１５４】

またさらに、上述のようにして判別した文書タイプを元に、特定ＵＲＬの文書データの中から特定の文書タイプの文書データを選択し、これら選択された文書データについての上述のリンク重要に基づいてリンク集を作成することにより、より質の高い特定分野の文書データにアクセスすることを可能とするリンク集を作成することが可能となる。

【図面の簡単な説明】

【図１】

第１実施形態に係わる文書検索装置の構成図である。

【図２】

文書情報を格納するテーブル集を示す図である。

【図３】

キーワード情報を格納するテーブル集を示す図である。

【図４】

索引情報を格納するテーブル集を示す図である。

【図５】

アクセスログを示す図である。

【図６】

索引生成処理の手順を示すフローチャートである。

【図７】

リンク重要度付与器において行う計算を模式的に示す図（その１）である。

【図８】

リンク重要度付与器において行う計算を模式的に示す図（その２）である。

【図９】

リンク重要度を算出する際にＵＲＬ類似度の概念を導入した結果を示す図である。

【図１０】

キーワード文字列グラフの一例を示す図である。

【図１１】

初期キーワード文字列グラフの生成手順を示すフローチャートである。

【図１２】

初期キーワード文字列グラフの生成手順を実現するアルゴリズムの一例を示す図である。

【図１３】

中間ノードの縮退処理の手順を示すフローチャートである。

【図１４】

中間ノードの縮退処理の手順を実現するアルゴリズムの一例を示す図である。

【図１５】

末端ノードの縮退処理の手順を示すフローチャートである。

【図１６】

末端ノードの縮退処理の手順を実現するアルゴリズムの一例を示す図である。

【図１７】

末端ノードの縮退処理をしたキーワード文字列グラフの一例を示す図である。

【図１８】

索引画面の遷移を示す図である。

【図１９】

索引トップ画面の一例を示す図である。

【図２０】

索引トップ画面のさらなる例を示す図である。

【図２１】

索引中間画面の一例を示す図である。

【図２２】

索引中間画面のさらなる例を示す図である。

【図２３】

索引中間画面のさらなる例を示す図である。

【図２４】

キーワード情報画面の一例を示す図である。

【図２５】

キーワード情報画面のさらなる例を示す図である。

【図２６】

第２実施形態に係わる文書検索装置の構成図である。

【図２７】

第３実施形態に係わる文書検索装置の構成図である。

【図２８】

第４実施形態に係わるリンク集生成システムの構成図である。

【図２９】

第５実施形態に係わるリンク集生成システムの構成図である。

【図３０】

情報処理装置の構成図である。

【図３１】

記録媒体を示す図である。

【図３２】

自明な読み入力インターフェースの一例を示す図である。

【符号の説明】

１１、１０２　処理装置

１２、１２３　入力装置

１３　表示装置

２１　リンク重要度付与器

２２　キーワード抽出器

２３　キーワード文書関連度計算器

２４　索引生成器

２５　索引アクセス部

２６　アクセス解析器

２７　ＵＲＬ類似度計算器

３０　文書データ

３１　リンク重要度

３２　ページキーワード

３３　索引データ

３４　アクセスログ

４１　文書情報テーブル

４２　被参照文書テーブル

５１　キーワードテーブル

５２　キーワード対応テーブル

５３　出現文書テーブル

６１　索引情報テーブル

６２　関連文書テーブル

６３　関連キーワードテーブル

７１　アクセスログ

８１、１０１　収集ロボット

８２　同義語データ

８３、１０７　ＷＷＷブラウザ

９１、１１１　文書タイプ判別器

９２、１１２　文書タイプ

１０３　ＵＲＬ文字列判別器

１０４　特定ＵＲＬ

１０５　リンク集

１０６　ＷＷＷサーバ

１２１　ＣＰＵ

１２２　メモリ

１２４　出力装置

１２５　外部記憶装置

１２６　媒体駆動装置

１２７　バス

１２８　ネットワーク接続装置

１２９　可搬記録媒体

１３０　プログラムデータ

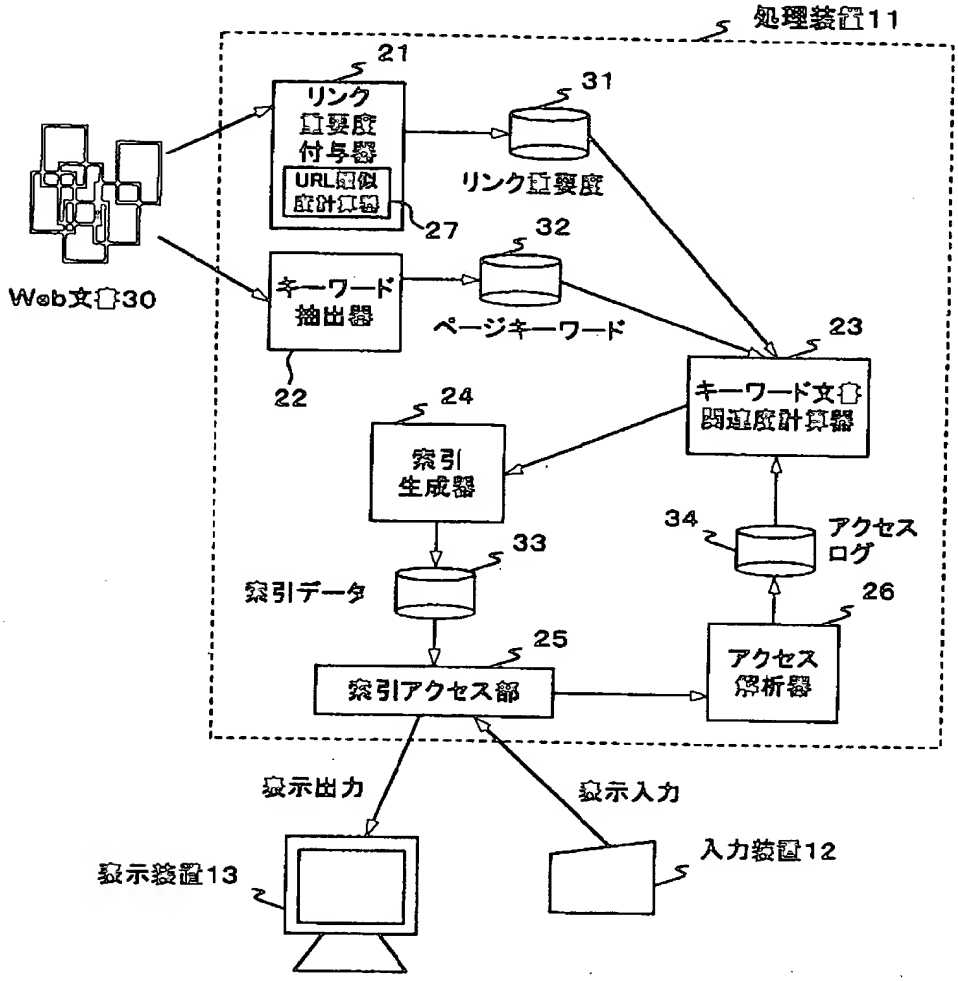
p、p_1、p_2、p_3、q、r_1、r_2、s_1、s_2　ページ

S　ステップ

【書類名】　　図面

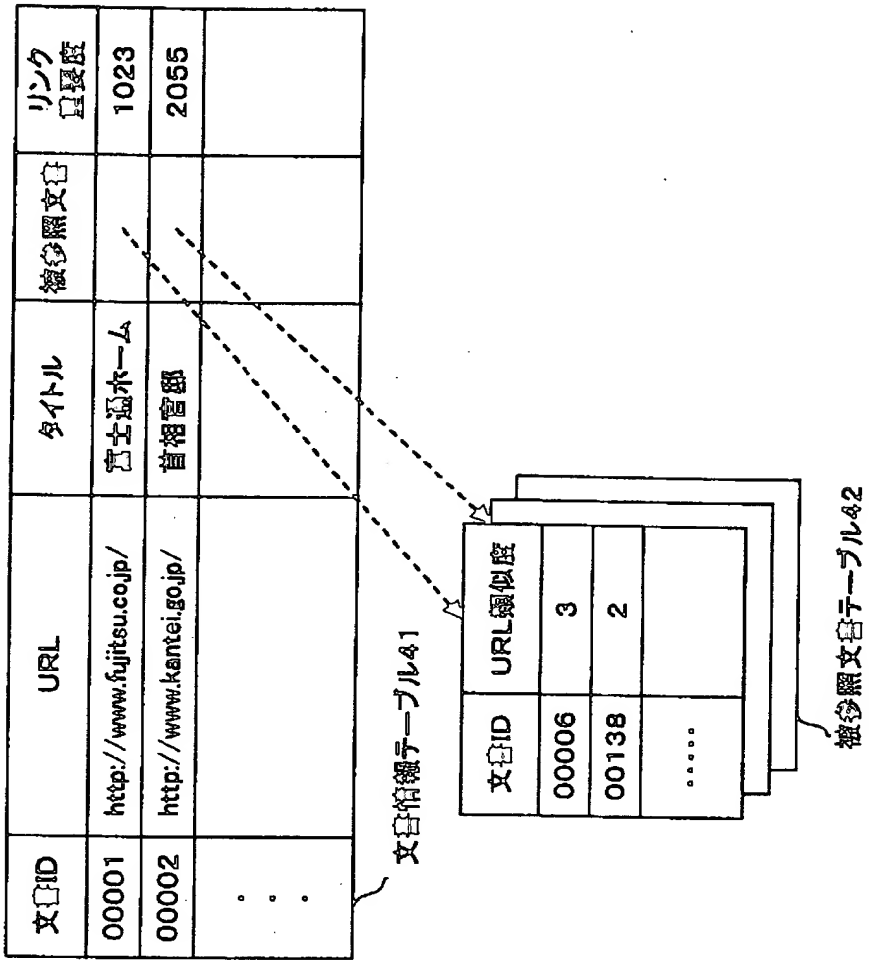
【図１】

第1実施形態に係わる文書検索装置の構成図



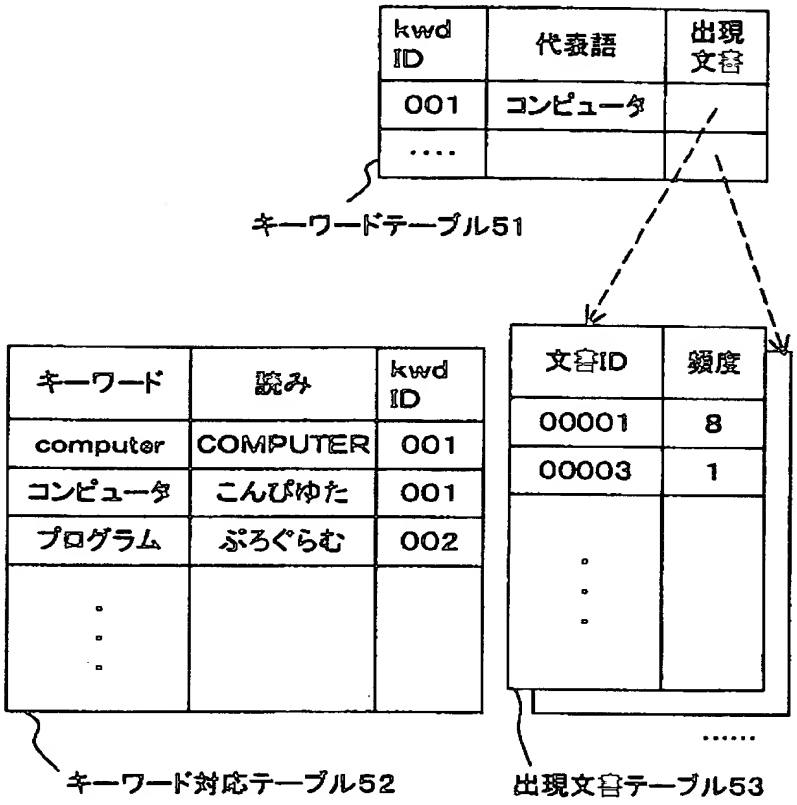
【図２】

文書情報を格納するテーブル集を示す図



【図３】

キーワード情報を格納するテーブル集を示す図



【図４】

索引情報を格納するテーブル集を示す図

文字列	継続する文字列	キーワード列
top	あ, い, ...	
あ	あいぼ, あお, ...	
あいぼ		相棒,アイボリー,
あお	あおぞ	青, 蒼, ...

索引情報テーブル61

キーワードID	関連文書ID列
093 321	0005, 0008, .. 0004, 0008, ...

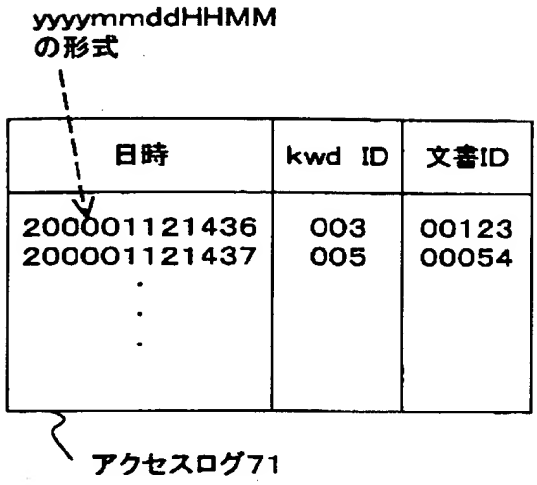
関連文書テーブル62

文書ID	関連キーワードID列
0005 0008 . . .	093, 099, 122, ... 093, 156, 321, ...

関連キーワードテーブル63

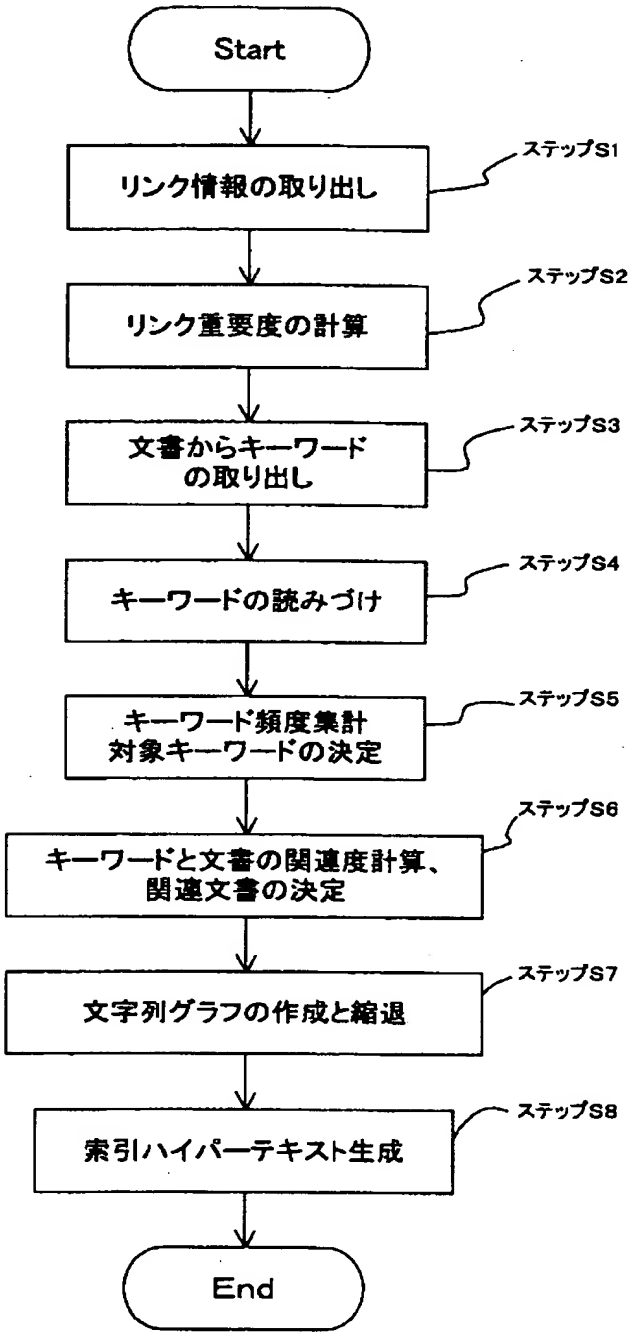
【図５】

アクセスログを示す図



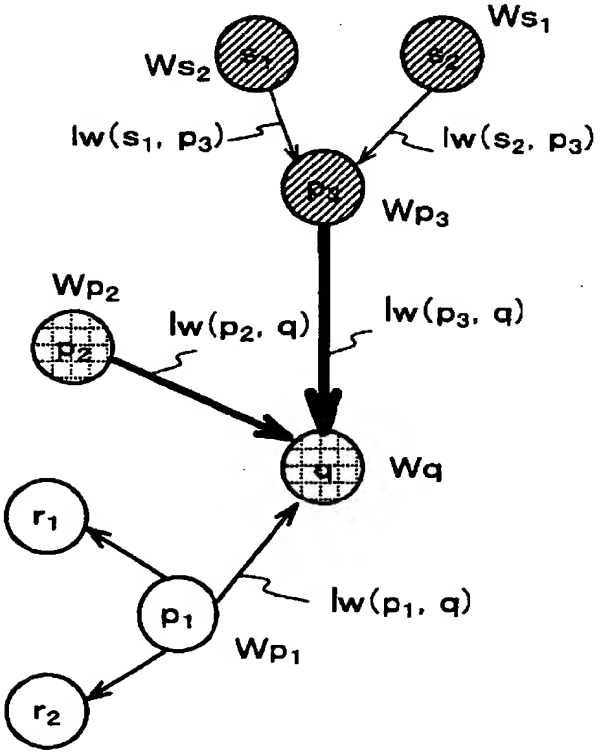
【図６】

検索処理の手順を示すフローチャート



【図７】

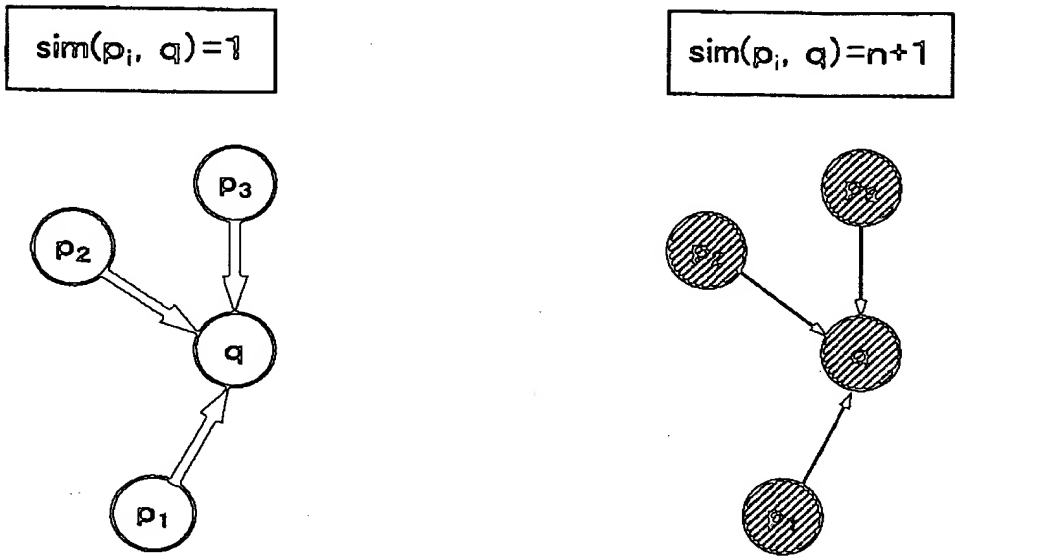
リンク重要度付与器において行う計算を
模式的に示す図（その１）



○印：wwwページを示す
○印の太さ：リンクの重みを示す
矢印の太さ：リンクの重みを示す
////：URLの類似

【図８】

リンク重要度付与器において行う計算を模式的に示す図（その2）



$$lw(p_i, q)=\frac{1}{sim(p_i, q)}=1$$

$$w_q=c_q+w_{p1}+w_{p2}+w_{p3}$$

図8(a)

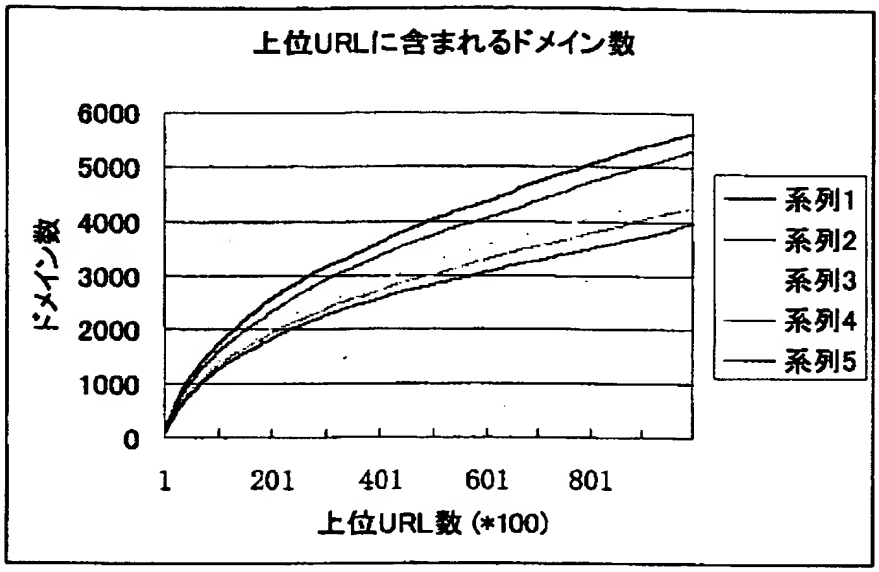
$$lw(p_i, q)=\frac{1}{sim(p_i, q)}=\frac{1}{n+1}$$

$$w_q=c_q+\frac{w_{p1}+w_{p2}+w_{p3}}{n+1}$$

図8(b)

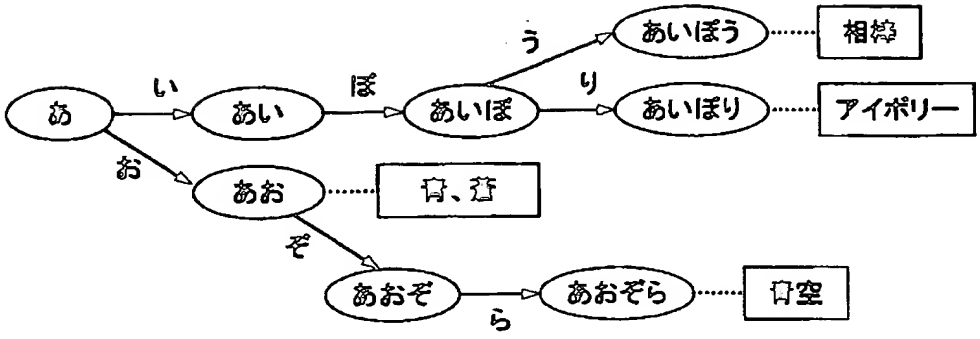
【図９】

リンク重要度を算出する際に
URL類似度の概念を導入した結果を示す図

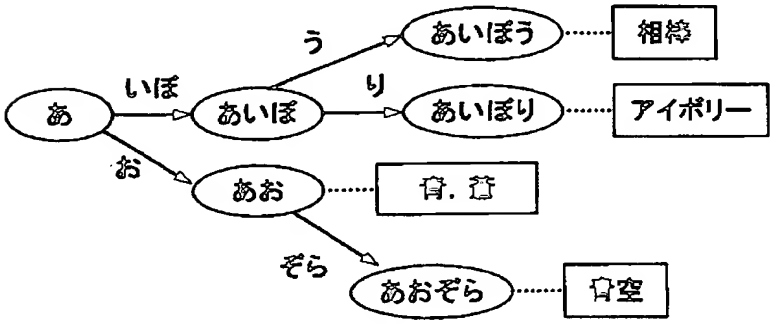


【図１０】

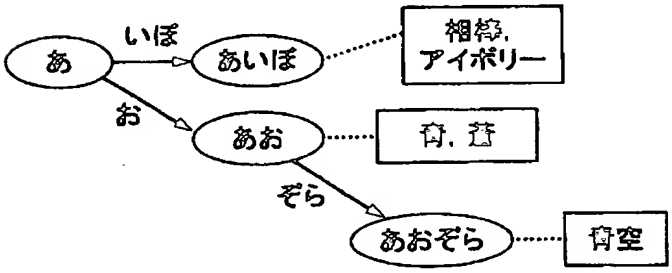
キーワード文字列グラフの一例を示す図



10(a)：初期キーワード文字列グラフ



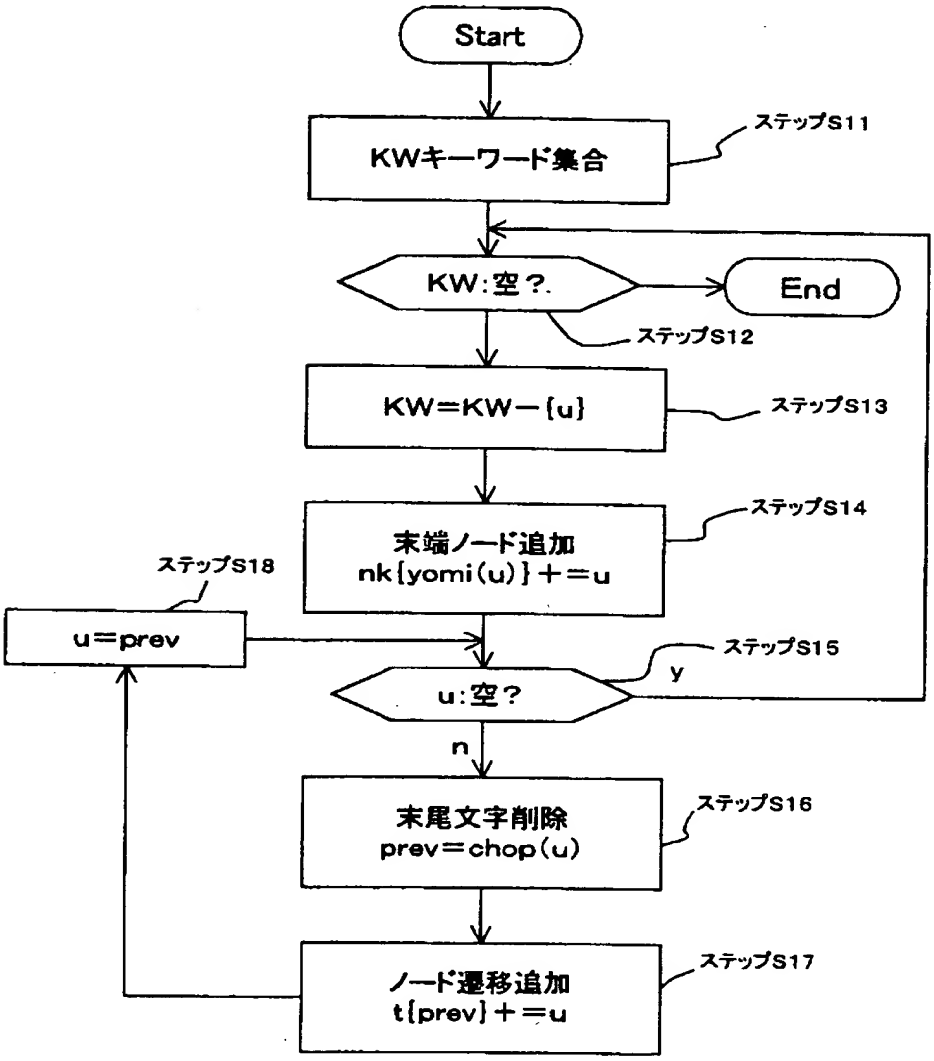
10(b)：中間パスの縮退後のグラフ



10(c)：末端ノードキーワードの縮退後のグラフ

【図１１】

初期キーワード文字列グラフの
生成手順を示すフローチャート



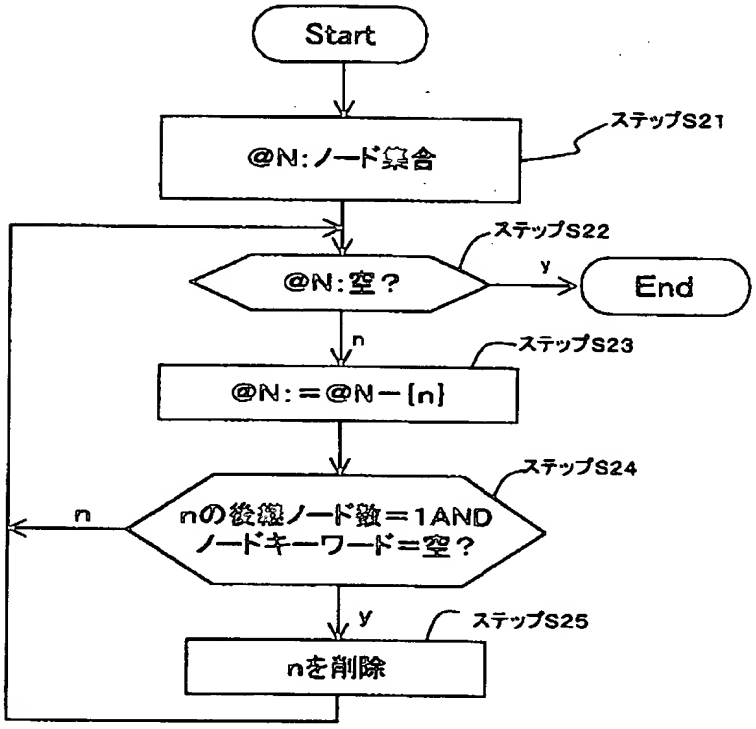
【図１２】

初期キーワード文字列グラフ生成手順を
実現するアルゴリズムの一例を示す図

```
proc init_kw_graph ()
{
  @KW:set of keywords;      # キーワード集合
  yomi : YOMI/Spell of keywords; # キーワードの読みを返す関数or 配列
  foreach u in KW {          # 各キーワードuについて
    nk{yomi{u}} .= u."+";    # u の読みノードの nk( ) を設定
    for ( i=0; i<length(u); i++) {   # u の文字列長ぶん繰り返し
       local prev = chop(u);          # u の末尾を切取ると親ノード
       t{prev} .= u."+";
       u = prev;
    }
  }
}
```

【図１３】

中間ノードの縮退処理の手順を示すフローチャート



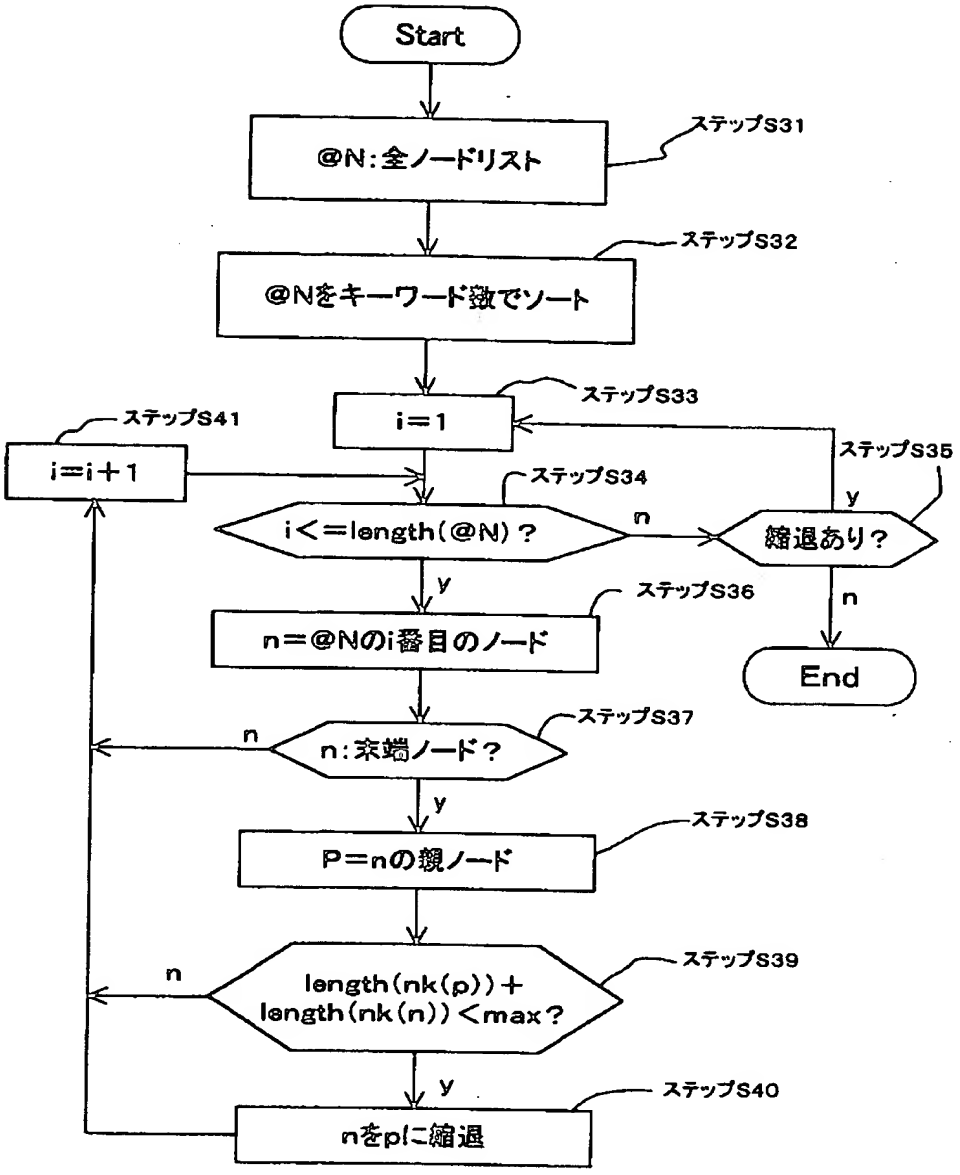
【図１４】

中間ノードの縮退処理の手順を実現する
アルゴリズムの一例を示す図

```
proc shrink_middle ( )
{
  @N : set of nodes
  foreach n (@N) {
    next = t{n} ;     # 次のノードリスト
    kw = nk{n} ;      # キーワードリスト
    if (length(next) ==1 && kw == "") {
       delete(n)          # ノードn削除
    }
  }
}
```

【図１５】

端末ノードの縮退処理の手順を示すフローチャート



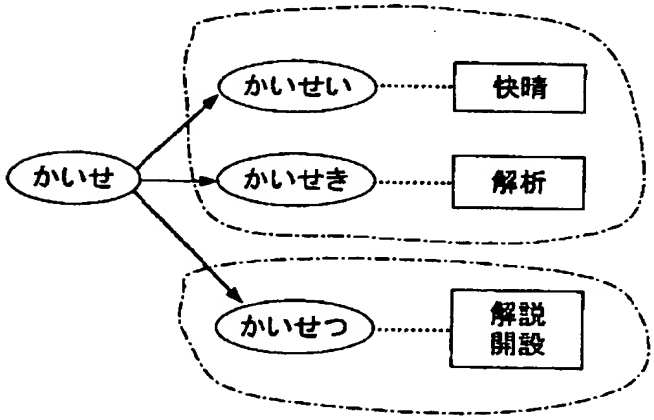
【図１６】

末端ノードの縮退処理の手順を
実現するアルゴリズムの一例を示す図

```
proc shrink_leaf ()
{
   @N: set of nodes;      # ノードリスト
   word_max =2;              # word_max ここでは２
   changed = true;           # キーワード移動した場合true
   @N = sort by_nk_length @N;  # キーワード数の少ない順にソート
   while (changed) {         # 移動がある限り続ける
      changed = false;
      foreach n in @N {
      if (is_leaf(n)) {      # 末端ノードの場合
         p = parent_node(n);      # 親ノード
         if (length(nk{p}) + length(nk{n}) < word_max) {
            nk{p}. = nk{n}. "+"; # キーワード移動
            delete (n);     # 末端ノード削除
            changed = true;   # 移動した証
         }
        }
      }
   }
}
```

【図１７】

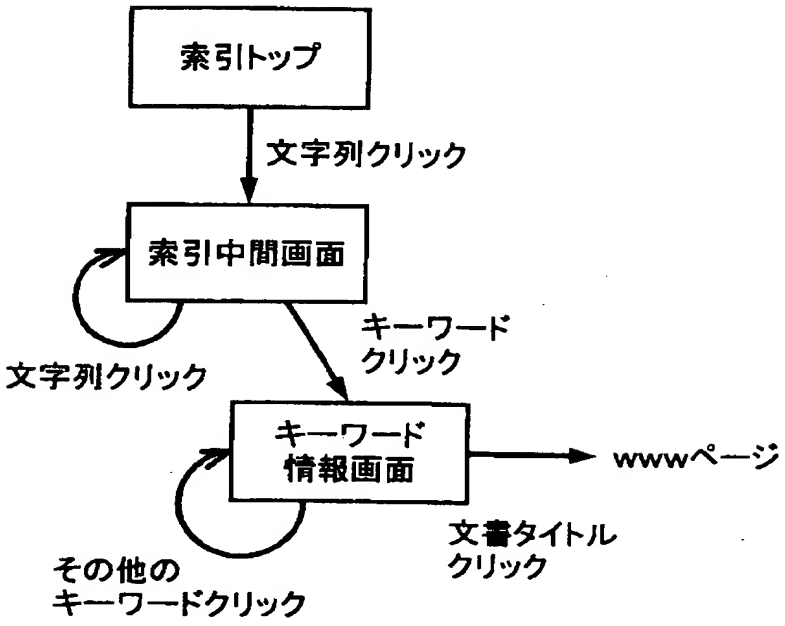
末端ノード縮退処理をした
キーワード文字列グラフの更なる一例を示す図



出証特２０００－３０７９６６７

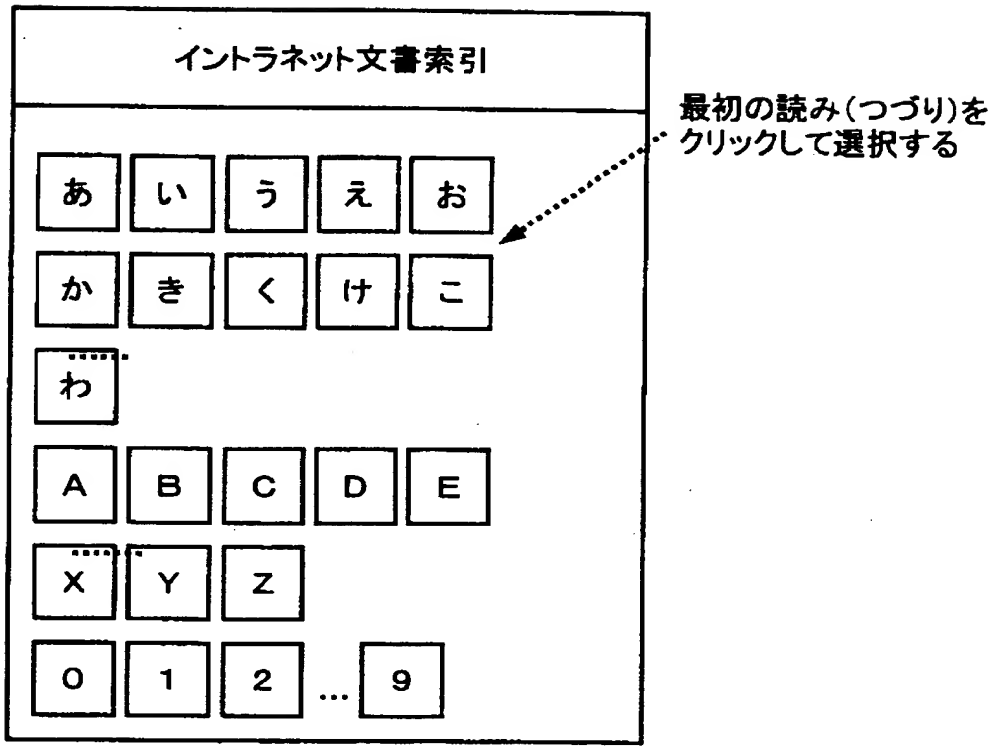
【図１８】

索引画面の遷移を示す図



【図１９】

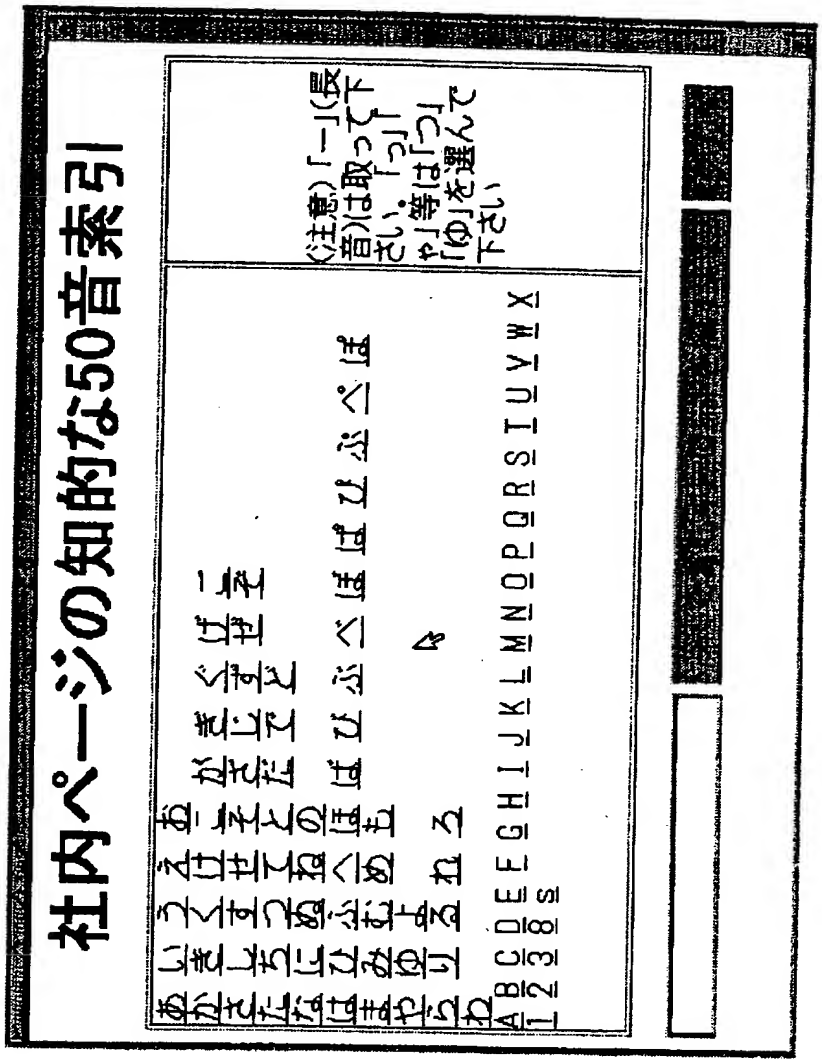
索引トップ画面の一例を示す図



索引トップ画面

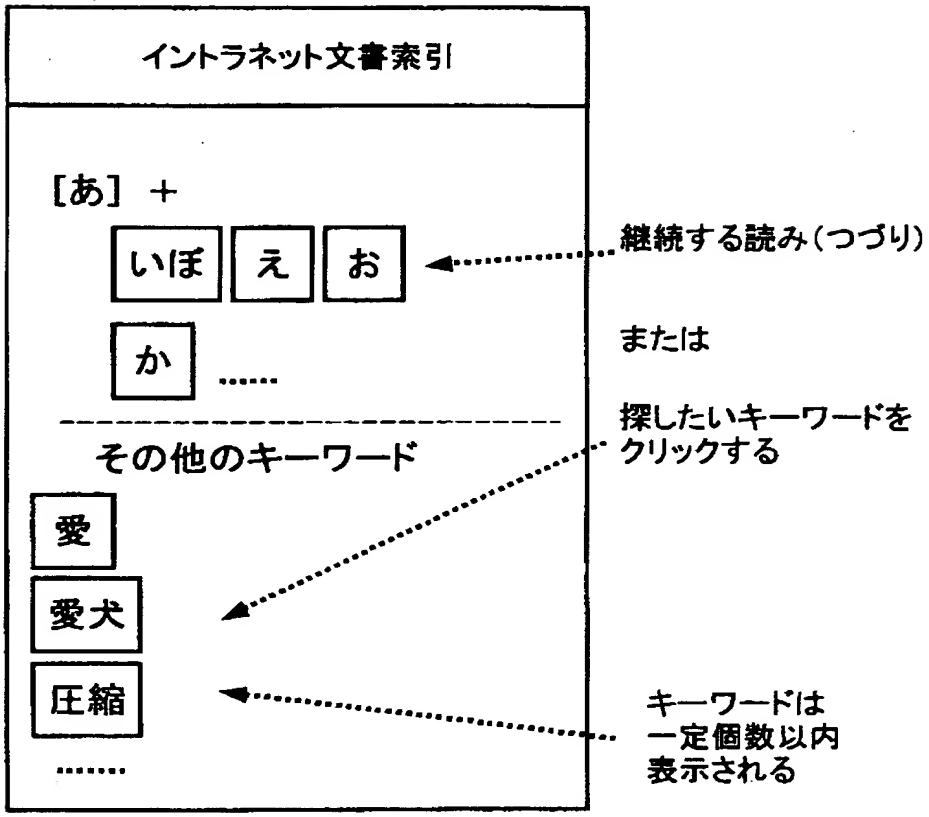
【図２０】

索引トップ画面のさらなる例を示す図



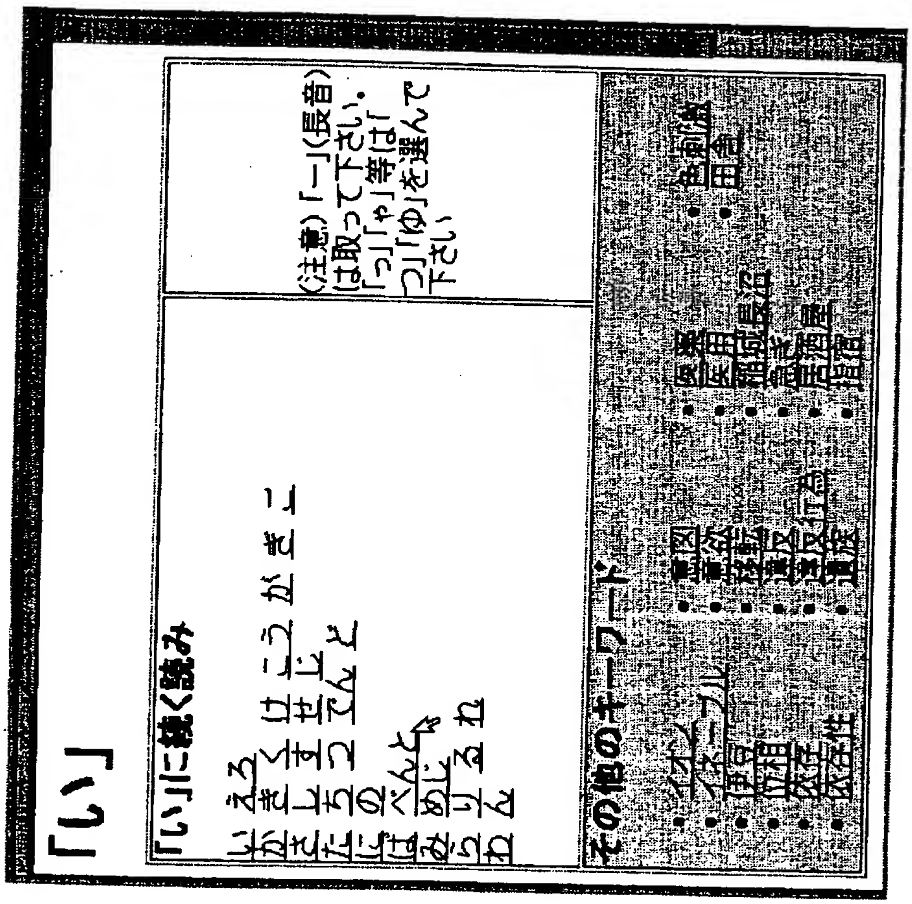
【図２１】

索引中間画面の一例を示す図



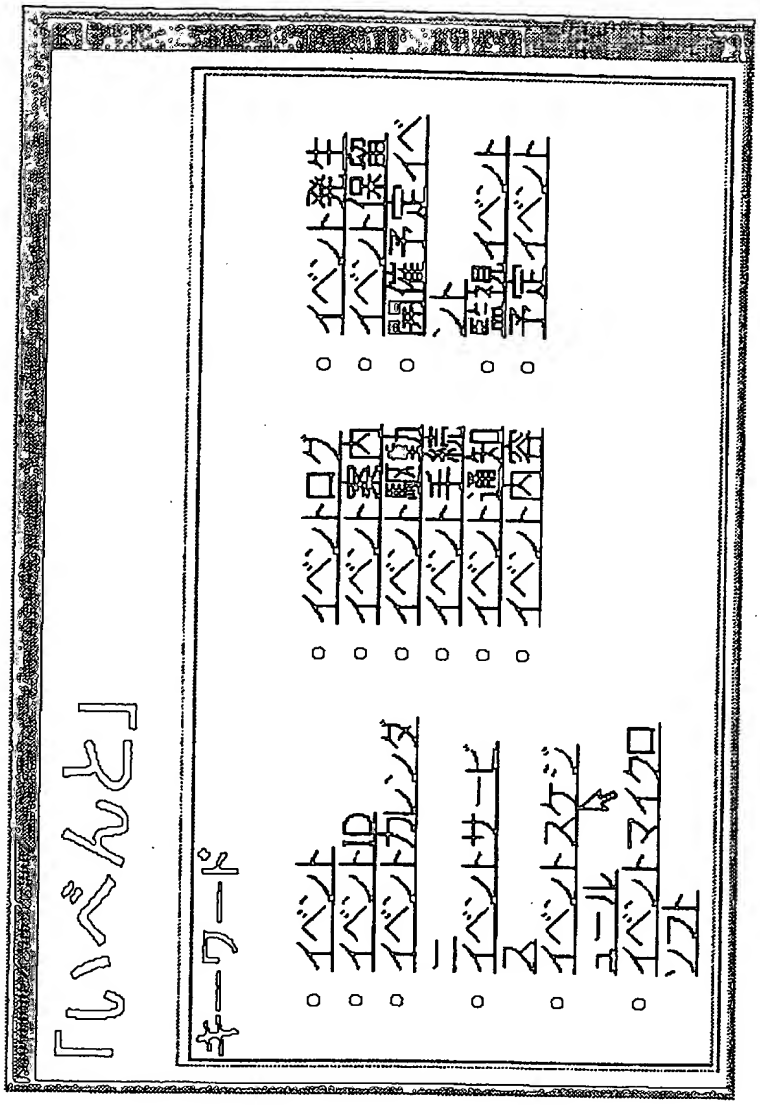
【図２２】

索引中間画面のさらなる例を示す図



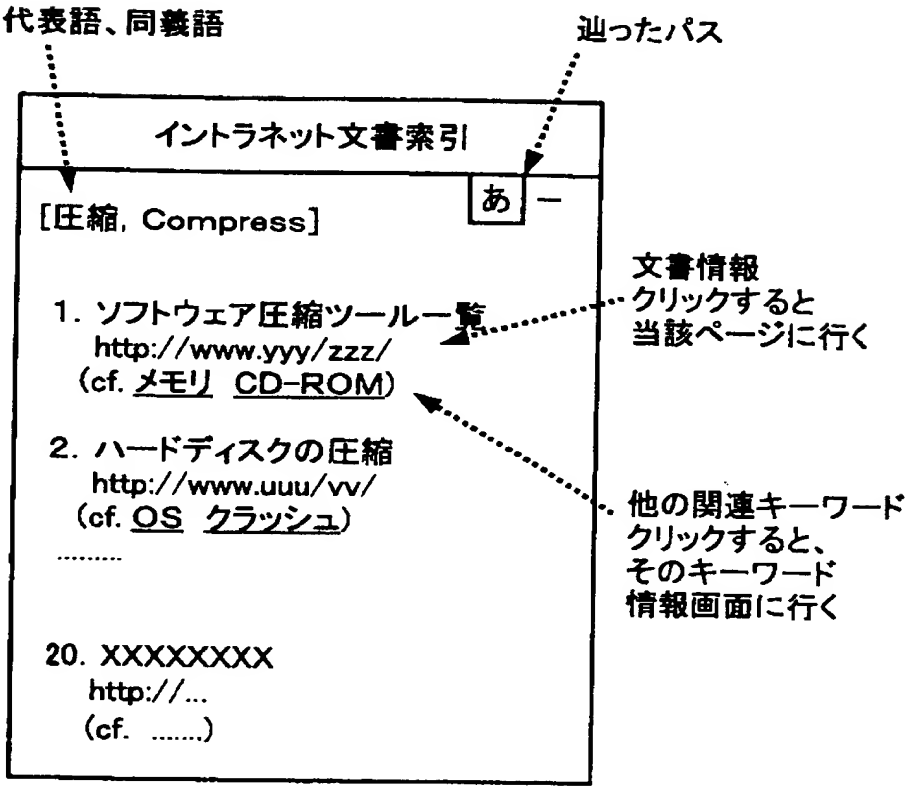
【図２３】

索引中間画面のさらなる例を示す図



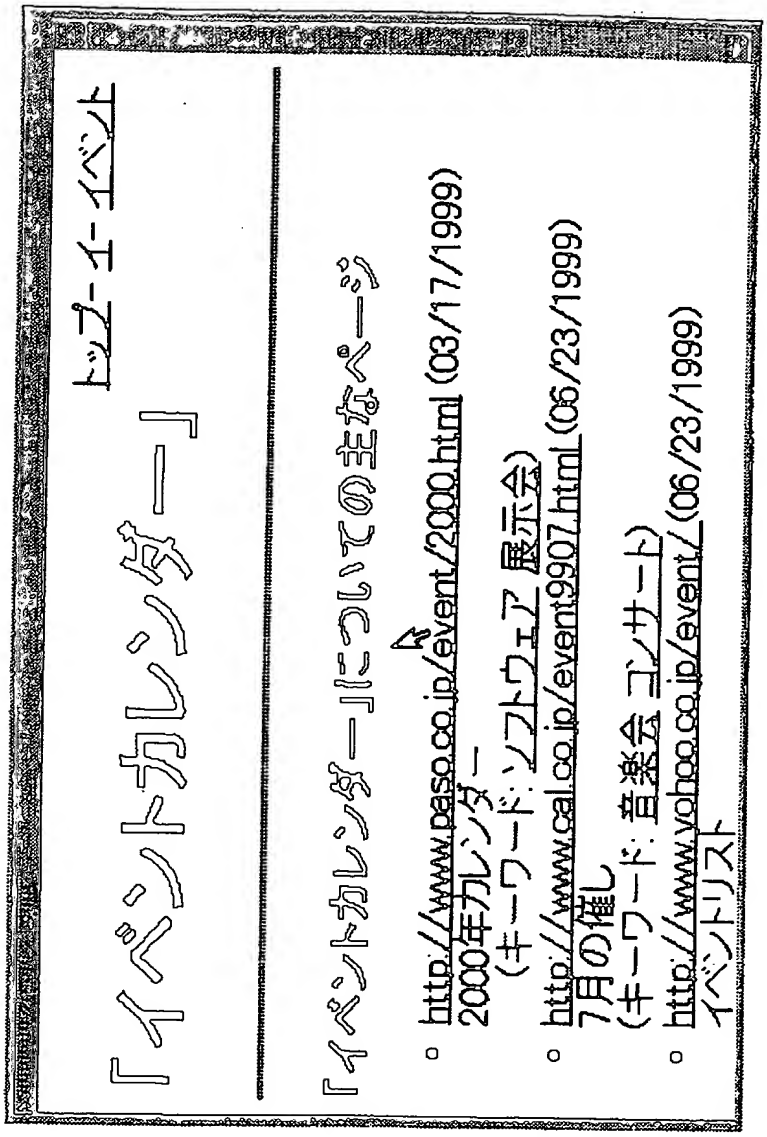
【図２４】

キーワード情報画面の一例を示す図



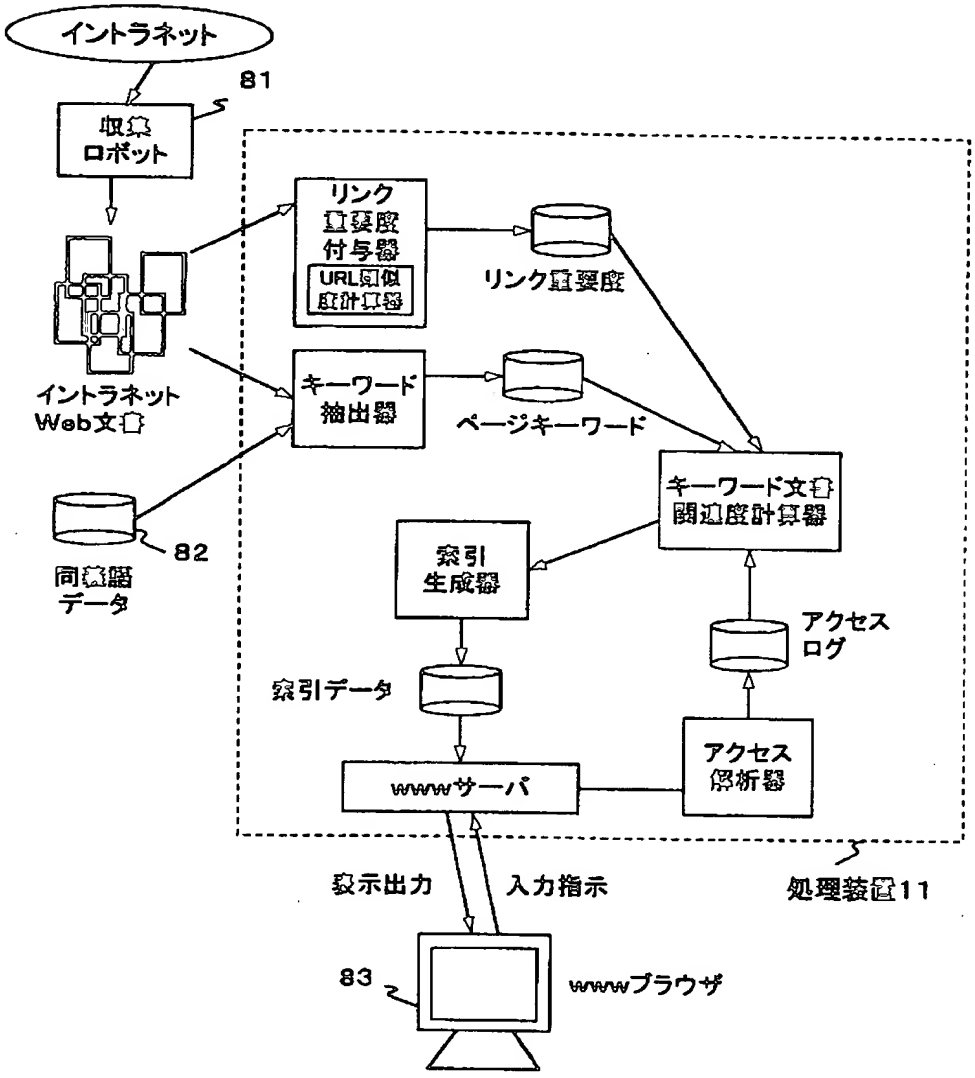
【図２５】

キーワード情報画面のさらなる例を示す図



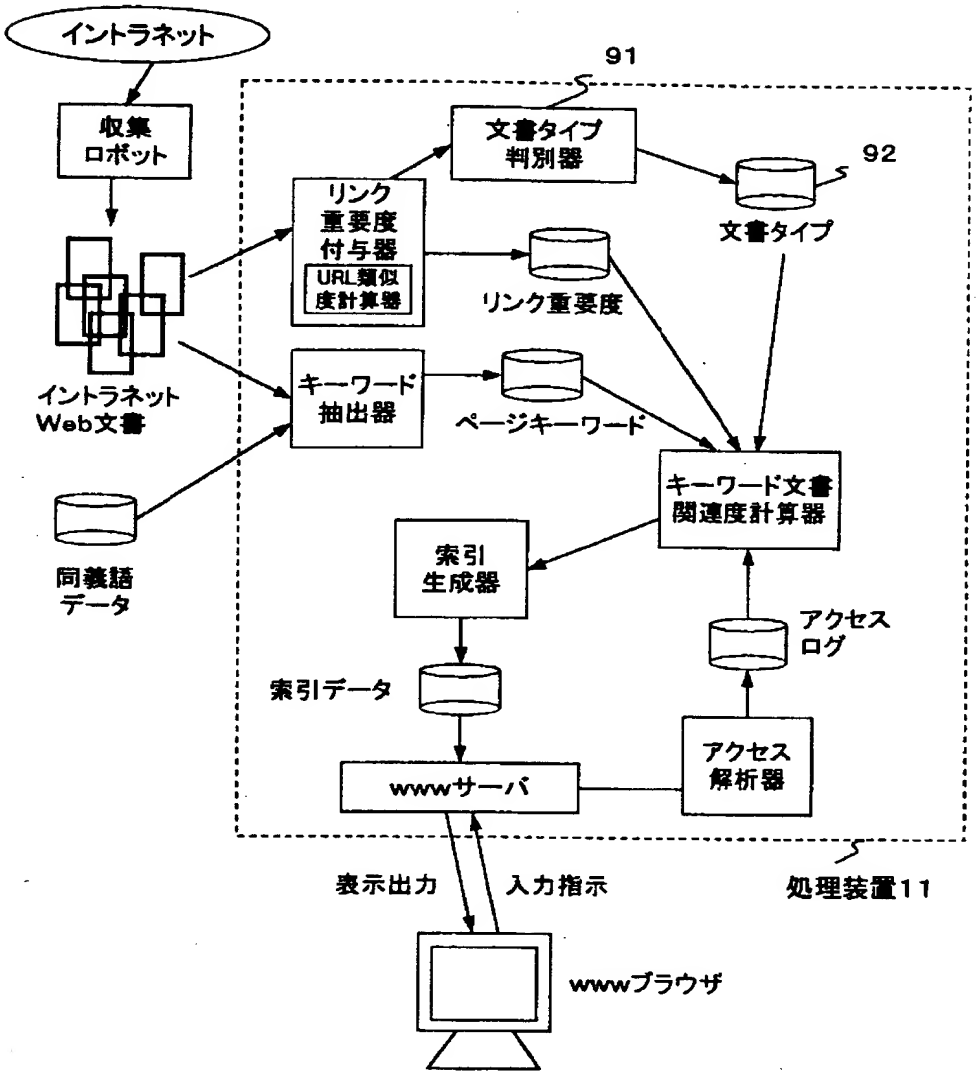
【図２６】

第2実施形態に係わる文書検索装置の構成図



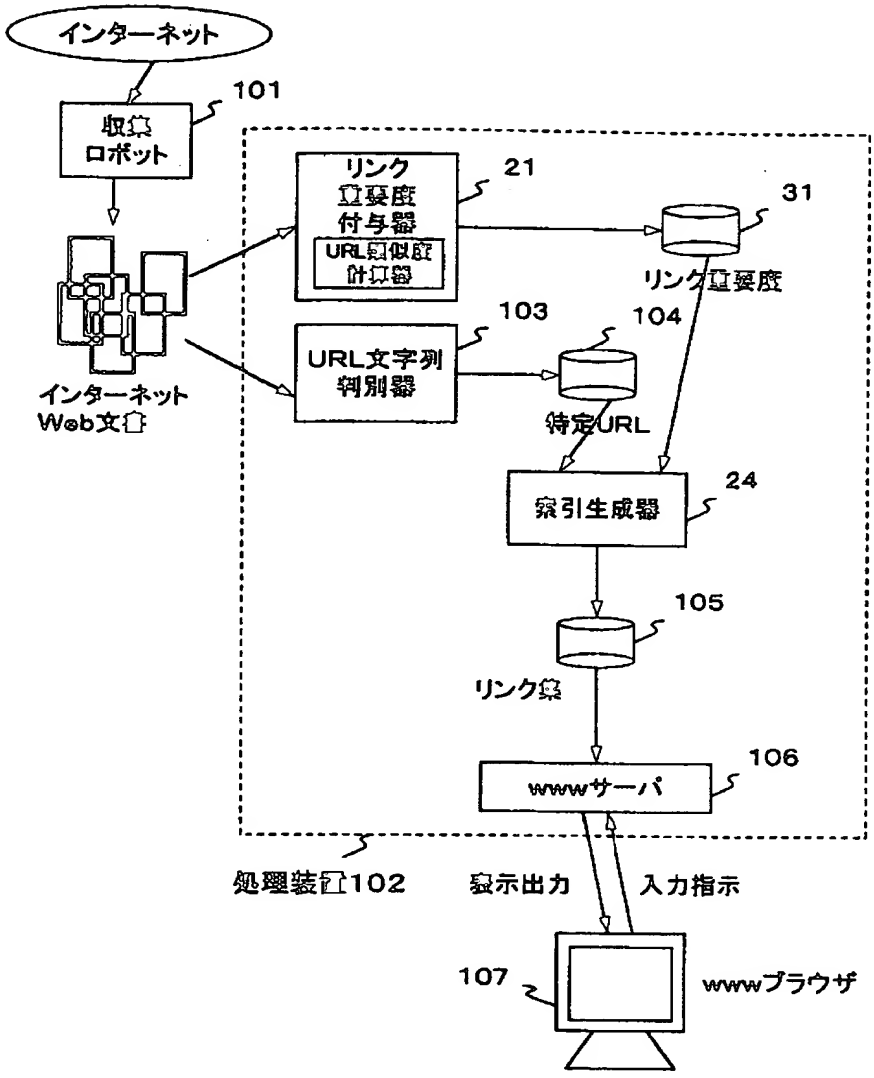
【図２７】

第３実施形態に係わる文書検索装置の構成図



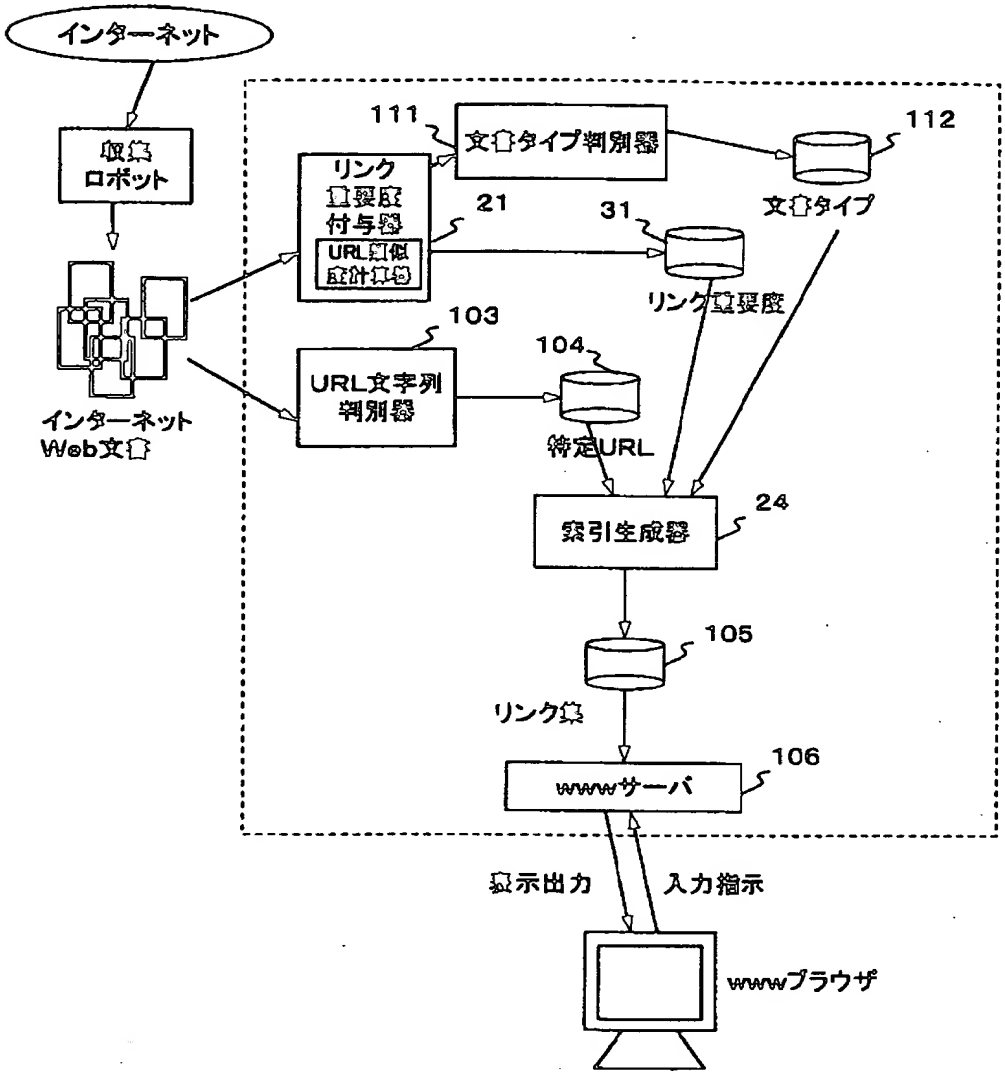
【図２８】

第４実施形態に係わるリンク集生成システムの構成図



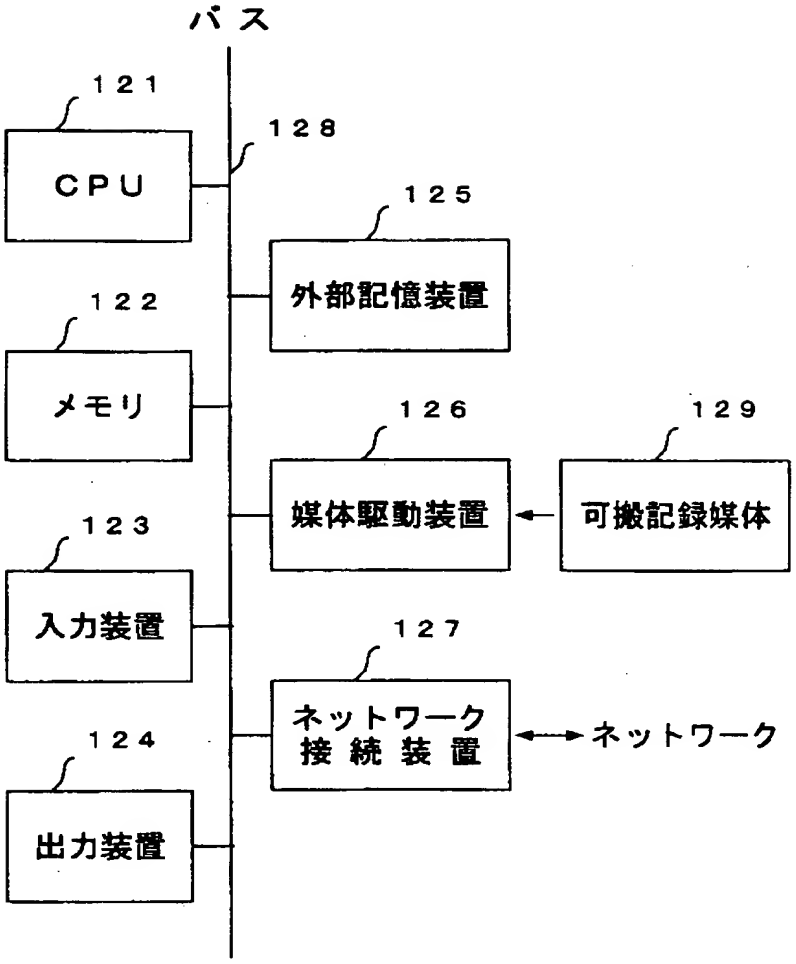
【図２９】

第５実施形態に係わるリンク集生成システムの構成図



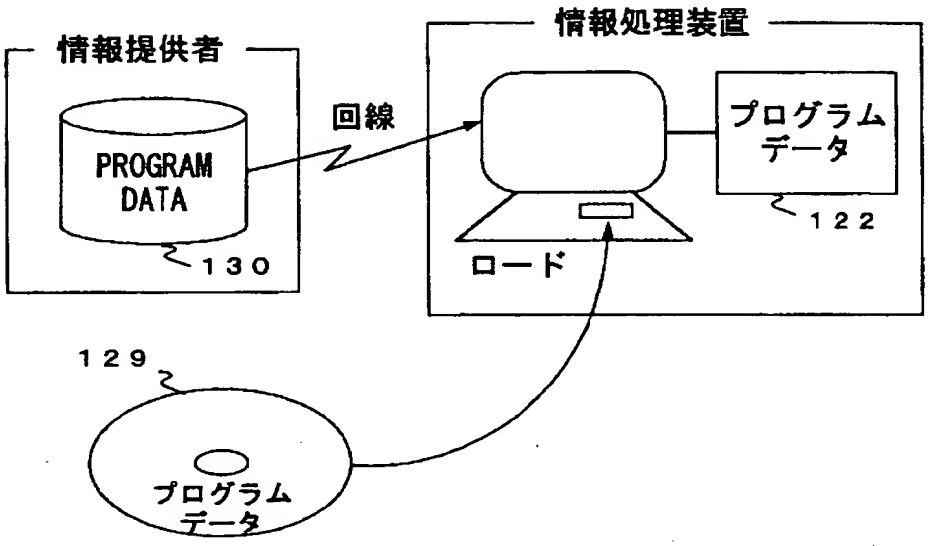
【図３０】

情　報　処　理　装　置　の　構　成　図



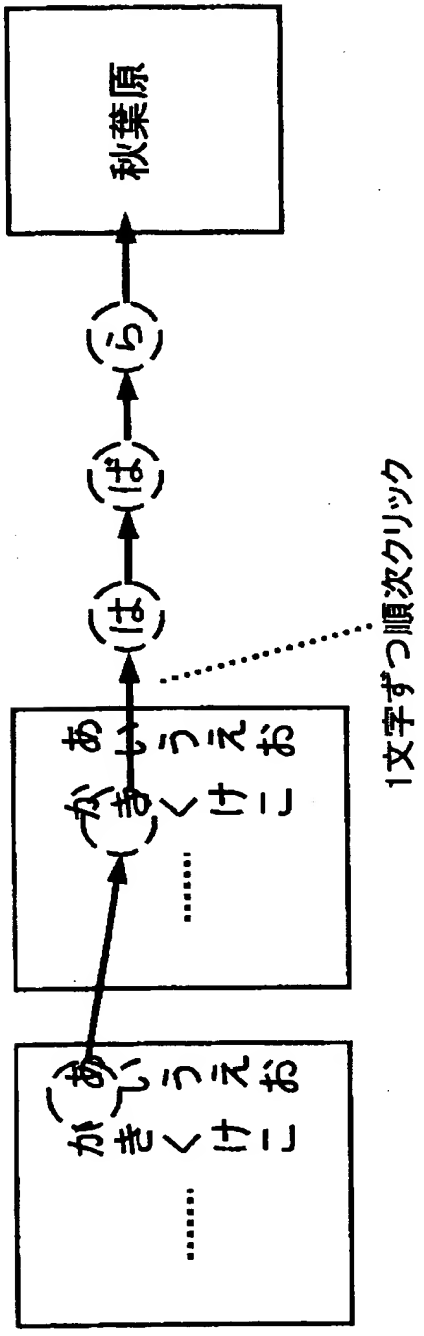
【図３１】

記　録　媒　体　を　示　す　図



【図３２】

自明な読み入力インタフェースの
一例を示す図



【書類名】　要約書

【要約】

【課題】　重要な文書データを自動的に検索することを可能とする。

【解決手段】　リンク関係を有する文書データ群から文書データを検索する文書検索装置にリンク重要度付与器２１及びアクセス部２５を備える。リンク重要度付与器２１は文書データ３０のリンク関係に重みを付けて、重要度であるリンク重要度３１を文書データ３０に付与し、アクセス部２５はリンク重要度３１に基づいて文書データ３０にアクセスする。

【選択図】　図１

出証特２０００－３０７９６６７

出　願　人　履　歴　情　報

識別番号　　〔０００００５２２３〕

１．変更年月日　　１９９６年　３月２６日
［変更理由］　住所変更
住　所　神奈川県川崎市中原区上小田中４丁目１番１号
氏　名　富士通株式会社